Departament of Signal Theory and Communications

University Carlos III of Madrid

Engineering Degree in Telecommunications Technologies

Bachelor Thesis

AUTOMATIC SELECTION OF FEATURES IN MRI FOR ALZHEIMER DETECTION

Author:

Nuño de la Concha Vega

Tutor:

Dr. Jesús Fernández Bes

February 2016

Acknowledgements

My family, for making me feel every day that I am a privileged.

Dr. Vanessa Gómez Verdejo, and my tutor, Dr. Jesús Fernández Bes, for helping me. Help based on their patience, their time and their limitless knowledge.

The possibility of doing this project, for teaching me to be honest with my work, for waking up my instinct in the right moment, in the challenges moment. Not forget the working nights, where hiding was not an option.

Despite the rain, I went for a run. I wanted to clean the mind and get wet. I ran with my thoughts. I stopped at a red light. Something caught my attention. Beside me, a man and a woman, both covering themselves with an umbrella. She, young. He, an old man. His white and curly hair, his short stature. His blue eyes and his relaxed face. I recognized him. I had seen several times his TV program. He was a very clever and a very curious man. I was glad to see him. I returned to my thoughts. And I was taken back from them again. "Are you cold?", she asked the old man. I was confused. Confused until I saw that the old man was barefoot. The man, the old and clever man, refused with a smile. He was lost. Lost in that way where everybody knows it except you. Lost in the reality of the moment. Moment when you lose everything you have learned, forget everything lived. You stop being yourself.

Abstract

The Alzheimer is such a serious and damaging disease for cognitive functions and mental abilities, as cruel to patients and their families when it comes to live and having to deal with the social and economic implications involved.

The increase in the number of patients in advance age as well as its mortality, in recent years has induced a greater awareness in society generating research projects to get an early diagnosis to decrease the effects of the disease, and who knows, if in a future, a cure. This early diagnosis, besides the positive impact for the health of the patients, would make a high cost reduction.

This project implements machine learning approaches for an automatic selection of features in MRI brain scans to detect Alzheimer.

For the project development, it is primarily done a study of the state of the art in the use of machine learning for mental disorders. It goes deep into the methods of selection of variables and linear classifiers appropriated for an optimal classification between Alzheimer and Control subjects, seeking the greatest possible efficiency.

In parallel, we analyze and download, for subsequent preprocessing, the images acquired at the Alzheimer's Disease Neuroimaging Initiative (ADNI) database available for medical research in Alzheimer.

Finally are implemented the variable selection algorithms, which provide the most relevant variables and being visualized to observe them in the context of a brain, being able to observe the areas with greatest impact for the detection of Alzheimer.

The results show a high percentage classification, these may be useful methods for optimal and early diagnosis, always having to be supervised by a medical specialist.

Resumen

El Alzheimer es una enfermedad tan grave y dañina para las funciones cognitivas y capacidades mentales, como cruel para los pacientes y sus familiares a la hora de convivir y tener que afrontar las repercusiones sociales y económicas que conlleva.

El incremento del número de enfermos en edad avanzada así como su mortalidad, ha tenido como consecuencia en los últimos años una mayor concienciación en la sociedad generando proyectos de investigación para conseguir un diagnóstico precoz que disminuya los efectos de la enfermedad, y quien sabe, si en un futuro, curarla. Este diagnóstico precoz, además del impacto positivo en la salud de los pacientes, lograría una reducción del coste económico de la enfermedad.

Este proyecto implementa procesos de lenguaje máquina para automatizar la selección de características en imágenes por resonancia magnética de cerebros para poder detectar el Alzheimer.

Para el desarrollo del proyecto se realiza en primer lugar un estudio del estado del arte en el uso de lenguaje máquina para trastornos mentales. Se profundiza en los métodos de selección de variables y clasificadores lineales más apropiados para realizar una óptima clasificación de los sujetos de Alzheimer y de Control, buscando la mayor eficiencia posible.

Paralelamente, se analizan y descargan, para luego preprocesarse, las imágenes adquiridas en la base de datos Alzheimer's Disease Neuroimaging Initiative (ADNI), disponible para investigaciones médicas en Alzheimer.

Finalmente se implementan los algoritmos de selección de variables, los cuales proporcionan las variables más relevantes y se visualizan para observarlas en el contexto de un cerebro, pudiendo observar las zonas con mayor impacto para la detección del Alzheimer.

Los resultados muestran unos altos porcentajes de clasificación, pudiendo resultar estos métodos muy útiles para lograr un diagnóstico óptimo y precoz, siempre teniendo que ser supervisados por un especialista.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem description

Alzheimer's disease (AD) is a neurodegenerative disease that appears in a human being by means of an aggressive and immediate cognitive deterioration and behavioral disorders of the superior brain functions (memory, language, orientation and spatial perception among others) due to the constant and persistent progress of the neurons death and the atrophy of certain brain areas. This deterioration leads to the loss of the autonomy of the patient, which becomes more dependent from his family and friends environment, having as a consequence a decrease and a negative impact in the social, work and fun activity of the patients and their caregiver responsibles [60][17]. Alzheimer's disease is the most common form of dementia for humans [17].

It is remarkable that some years ago the references to dementia in the medical literature where just a few, existing 3 in 1935, 25 in 1950 and then increasing to $90,000$ in 2007, as a consequence of an active research in the field of dementia [17]. It is obvious the increasing interest and awareness of the problem by the society. A justified reason for this social preoccupation has to be with the progressive aging of the world population and that Alzheimer is one of the main reasons of incapacity and dependency in people older than 65 years, with a high mortability, causing a huge economic, social and sanitary cost that mainly suffers the family [17].

Alzheimer is in third position of order of disease importance in the United States. In terms of costs for the society, in the year 2009 had an outgoing of $97,000$ millions of dollars. In Europe, for that same year, the cost was of $103,300$ millions of euros, and in both cases just for direct costs (medical attendance, complementary studies, medicaments, ...). In Spain, attending to estimations done in year 2004, the total cost for dementias, including public and private, was higher than $8,200$ millions euros [17].

Attending to predictive models, an early diagnostic can have a positive impact for the health of the patients, and could make a high cost reduction of 10% [35]. As a consequence, it is highly recommended the implementation of strategies to get an early diagnostic as a measure for cost-effectiveness [37].

To this end, neuroanatomical and neurofunctional assessments have increased and become a common practice for the diagnostic of a mental disorder pathology. The Alzheimer's disease diagnosis is achieved by two ways: the realization of cognitive and behavior eval-

uations and by neuroimage analysis [18]. The functional and behavioral assessments are carried out by the Clinical Dementia Rating (CDR) or the Geriatric Depression Scale (GDS), among others.

In the case of the neuroimaging evaluation, it has grown the interest in applying Multivariate Pattern Analysis (MVPA) to study the patterns of neurodegenerative diseases and mental disorders using Magnetic Resonance Imaging (MRI), structural (sMRI) and functional (fMRI) [20][23]. In the last years, diseases such as Alzheimer (AD), Huntington (HD), major depression disorders (MDD) or schizophrenia, among others, have been applied to different neurofunctional technics as Voxel-Based Morphometry (VBM), which is an analysis technique in neuroimaging that allows the researching of possible differences in the brain anatomy [11][12].

This evolution in the evaluation of some mental disorders by means of MRI, has promoted the research and the improvement of better neuroimaging analysis techniques and their methods and algorithms for using them. One of these improvements is the application of Machine Learning (ML) approaches to characterize patterns present in neuroimaging data, and justified by the fail of the translation of results of other models from the neuroanatomical hyphotesis to the clinical practice [55].

ML represents a reliable option for the extraction of relevant information from neuroimaging data by using statistical learning methods. The value of its performance has to be with the characteristic of automatically learning a model from a collection of samples, reaching some approaches to detect useful information that could not be detect by common and manual techniques, nor human eyes.

Besides, the capability of the ML to extract the highest amount of information from the available sample collection, even when this collection has a reduced and limited number of samples, has positioned ML as a principal tool for neuroimaging analysis [57].

1.2 Objectives of the project

All the methods and techniques performed in this project have been implemented with the principal objective of the automatic selection of features in MRI to detect Alzheimer. This objective can be divided into a group of subobjectives:

1. Selection of samples set of MRI brain scans with common imaging characteristics through the access and examination of the Alzheimer's Disease Neuroimaging Initiative 1.3 database. Following pre-established patterns of searching and downloading, we get as many Alzheimer and Control subjects as available in the database. Preprocessing of these MRI brain scans to obtain the features from the voxels of the MRI.

2. Application of feature selection techniques to reduce the high dimensionality of features on each MRI. Implementation of the RFE and t-test methods to eliminate redundant and irrelevant features, avoiding noise and difficult interpretation, and keeping the most relevant ones. Use of Synthetic data samples generated by us for a more clear understanding and easier development.

3. Use of a linear classification to distinguish between healthy and ill subjects. For reasons explained in Chapter 2, a linear classifier SVM will be implemented and

MATLAB toolboxes used for the visualization of the MRI and the development of the algorithms. Try to get the highest accuracy possible in each of the methods.

4. Evaluation of the performances of each of the implemented techniques. An analysis of the different strategies followed by each method and their classification error in the experiments. We are interested in getting a final visualization of the most relevant features in its context, a brain template.

1.3 Description of ADNI initiative

Created in October 2004, the Alzheimer's Disease Neuroimaging Initiative (ADNI) is a global research effort that "actively supports the investigation and development of treatments that try to slow or stop the progression of AD" [5]. Motivated by the huge impact of the Alzheimer in our society, affecting to almost the 50% of the people over the age of 85 years and in the top ten leading causes of USA deaths [6], the ADNI initiative has been validating and using clinical data as genetics, cognitive tests, blood biomarkers and MRI and PET[1] images. All this clinical data has been obtained from volunteers subjects from North America [5], from different ages (as shown in figure 1.1) and genders, and in all the different phases of the alzheimer's disease process: control volunteers (C), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) and dementia or Alzheimer disease (AD).

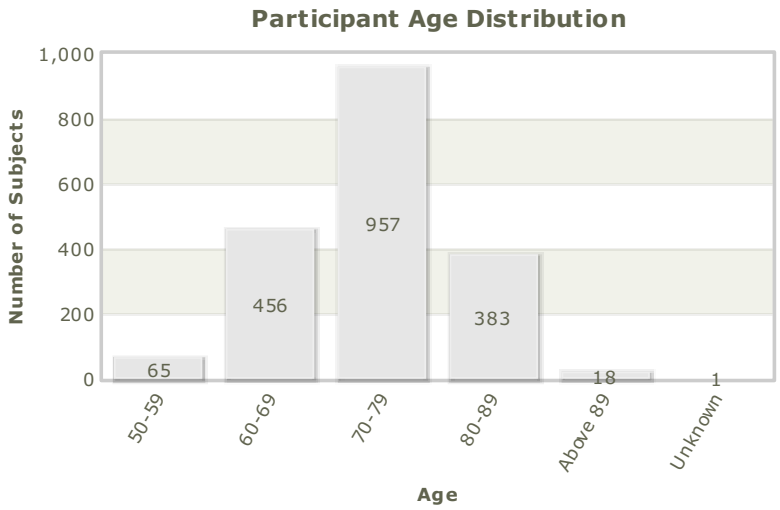


Figure 1.1: Age distribution ADNI participants. Graph extracted from ADNI official site.

Focusing on image data, which is what we have working with, ADNI provides an integrated and safe environment to archive neuroimaging data. A balance between allowing researchers to take advantage of the available resources and protection from unauthorized access is reached. This is possible thanks to an agreement and authorization in the user registration method, and the processes and patterns followed for the data acquisition, becoming a worldwide Alzheimer resources sharing [5].

Figure 1.2 shows the three phases followed in the ADNI study: ADNI1, ADNIGO and ADNI2. During each of these phases, its imaging and biomarker collection has been increased with new subjects. The medical condition of the subjects are followed over the

[1]Position Emission Tomography

time to have a control of the pathology disease progression. Nowadays the ADNI initiative counts with 10,454,426 downloads, 80,681 users and 595,033 uploads [5].

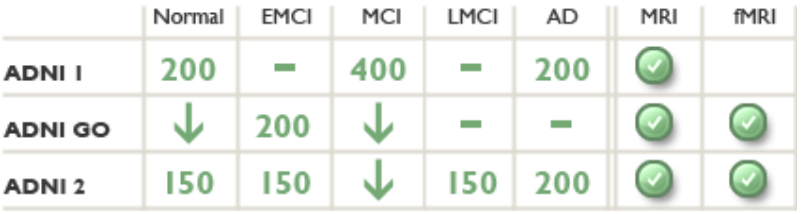


Figure 1.2: The table summarizes the North American ADNI study target participant numbers, as well as the types of data taken at the different phases. In reality, the total number of study participants vary. Graph extracted from ADNI official site.

1.4 Structure of the project

This bachelor thesis is organized in the following way. Chapter 2 provides a deep explanation of Machine Learning for the classification problem as well as the basic knowledge and formulas needed to understand the SVM linear classifier. At the end of this long chapter, feature selection techniques implemented in this project are described.

Chapter 3 begins with an introduction of the MRI image. A better understanding of its characteristics, MRI voxels definition and a graphical representation. Then, the process followed to acquire the data used during the experiments and its preprocessing process is explained. It goes into every detail to describe each step followed by the student in this phase.

The experiments implemented by the machine learning methods are explained in Chapter 4. Justifying the value of the parameters used in the model, visualization by figures of the accuracy of each methods, and graphical representation of the most relevant features.

In Chapter 5 are presented the conclusions obtained at the end of this project and an analysis of future lines of investigation to follow which are out of the scope of this bachelor thesis.

Finally, three appendices are included at the end of the project. In Appendix A is explained the planning of the project, with its different phases and the tasks implemented during the whole process and their duration. In Appendix B, the budget of the project is present, its personal and material costs. And the last one, Appendix , is a summary of the project.

Chapter 2

Machine Learning approximations to MRI processing

The existence of machine learning in computer science was motivated by the purpose of the development of algorithms that could learn from and make predictions or decisions based on data [15]. As a branch of the artificial intelligence, it is focused on reproducing generalized behaviors from training data inputs, generating an induction of the knowledge. As a consequence of reaching this possibility, a huge simplification of the scientific processes is obtained due to the partial automatization, drastically reducing the computational complexity/cost [53].

A simple definition to explain the concept of machine learning could be: "Field of study that gives computers the ability to learn without being explicitly programmed" [53], said by Arthur Samuel in 1959. Tom M. Mitchel was able to give a more formal and precise definition, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [53].

2.1 Introduction

One example of a machine learning problem is classification [15]. In such a problem, our objective is to predict the class of some data samples out of a discrete number of known classes. The inbox of the Google mail (Gmail) is an example of classification, where the emails are the inputs of the model and the classes in which are classified are the 3 labels, "Principal", "Social" and "Promotions".

In a first step in the learning process a large set of N inputs $\{x_1, ..., x_N\}$, called a training set, is used to set the parameters of an adaptive model by a Machine learning algorithm. This training set has been previously classified and labelled in their respective categories. The machine learning algorithm is run, which can be understood as having a function $y=f(x)$, where input x, training set, is used to generate an output y, labels. This process corresponds to the training or learning phase. To the model training follows the testing phase, where new data inputs, completely different to the ones corresponding to the training set, are used to measure the accuracy of the model generated. The ability of categorizing this new data set is known as generalization. This ability can be consider the main goal of this process since is needed to build a successful model that should be able

to classify different inputs, not just fit to the training ones, avoiding the risk of overfitting [15].

The aim of this project is to be able to identify the class of a subject, Alzheimer or Control, by means of their features. To the end of this application, the algorithm that has been used, produces a dependency between the inputs and the desired outputs of the system. This dependency is obtained through the learning acquired thanks to previous labelled samples training. This description corresponds to the algorithm of the supervised learning, one of the types of machine learning algorithms which follow a taxonomy in function of the generated outputs by the system [15].

This process of assigning each input vector to an output consisting on one of a finite number of discrete classes (binary or multiclass) is called classification. In case of looking for an output with one or more continuous variables, then it will be talking about a regression problem. Said a different way, Regression involves estimating or predicting a response. Classification is identifying a group membership. In this occasion a classification problem will be faced [15].

2.2 Classification

The classification of the subjects, based on the class they belong to, will be obtained by a linear classifier through a lineal combination of their characteristics. The characteristics, also known as features or even Neuromarkers, for this case, will be real numbers which corresponds to the voxels of the MRI images, as it has been previously mentioned in this project.

2.2.1 Linear Classifier

The linear classifiers are normally used in scenarios where the speed in classification is important, since most of the times is the fastest classifier. Frequently these classifiers perform quite well when the number of subject samples is big, like in our project [62].

Our study contains a data set with an extremely high number of dimensions, more than $500,000$ elements, representing the voxels of the MRI images, and a quite small number of subjects, more than 400. This means that our problem is separable; the ratio of possible decision thresholds is infinite and all of them corrects.

Since that, this simple lineal combination 2.1 can be used to expressed the lineal classifier, the parameters involved and the output generated,

$$\vec{y} = f(\vec{w}.\vec{x}) = f(\sum_{i=1}^{l} \mathbf{w_i x_i}) \tag{2.1}$$

Remembering that $\mathbf{x}$ is the input dataset, a vector NxD, being N the number of samples and D the dimension of the features of these samples. Then, N is the Alzheimer and Control patients and D the coefficients of the voxels. $\mathbf{w}$ is the weight's vectors and $f(.)$ is the function that converts the inner product two vectors in the desired output, $\mathbf{y}$. The weight vector $\mathbf{w}$ is learned from a training set. Often $f(.)$ is a simple non-linear function that maps all the values over a certain threshold to the first class and the others, to the second one, in case of a binary case, which is our case [62].

A huge advantage of this approach is that once the learning process using the training set has been completed, this training set can be discarded, since the parameter $\mathbf{w}$ has already learned and is what has to be kept [62].

The linear classifier is able to separate the samples by the class they belong to. Although there are multiple algorithms that solve the required problem, like the Perceptron or Fisher's classifiers, the Support Vectors Machines classifier will be the one used in this work, since our main objective is not to classify but to make a selection of the most relevant features of the MRI images, the ones which provide the highest level of information to classify subjects. The SVM will assign to each feature a weight according to its relevance.

2.2.2 SVM

The Support Vector Machines (SVM), created by Vladimir Vapnik and AT&T labs, are a family of supervised machine learning algorithms that are applicable to classification and regression problems [16][25][45].

SVM are characterized by their great potential in classification tasks, mainly by their excellent generalization capacity due to the fact that they are grounded in the theory of statistical learning [52][24][39][38].

There are just a few parameters which need to be set up; the model just depends on the data with the highest level of information [21]. Support Vectors Machines stand out not just but their capacity of generalizing the model, but also because of their capability for minimizing overfitting in the training set, capability that is not presented in other learning methods like the neural networks [15].

Support Vector Classifiers

In a binary classification problem, given a set of labeled training samples, an SVM can be trained to build a linear model that predicts the class of any new data sample. Intuitively, an SVM classifier is represented as a hyperplane that splits the input space into two separate regions, each belonging to one class. From all the possible hyperplanes that classify the input data, the SVM algorithm selects the one that maximizes the distance to the input data [13].

Although the SVM is fundamentally a linear classifier, it can be used to represent nonlinear ones by the uses of Kernel functions [15]. In this work, due to the high number of features with respect to the number of data points, the problem is linearly separable so we will focus on the standard linear SVM classifier, which we will derive in the following subsections.

Specifying, a SVM model builds up a hyperplane or a set of hyperplanes in a space of high dimensionality (even infinite). Infinite discriminatory hyperplanes, as illustrated in figure 2.1, exist to divide the samples, because of this, is desired the one which classify in the most optimum form. Figure 2.2 shows perfectly what this this means, to look for the highest margin of separation between samples of different classes, reason of why the SVM are also known as "classifiers of maximum margin" [21]. Following this, samples labelled in one class will be on one side of the hyperplane, and samples belonging to the other class, in the opposite side [13].

Classification by SVM can be carried out by a linear or non-linear separation in the surface of the samples, always with the main objective of minimizing the error classification, reaching for the highest accuracy [52].

The solution obtained from the SVM is *sparse*, this means that the majority of variables are zero in the solution of the model, letting this final model be able to be written as a combination of a small number of input vectors, called **Support Vectors**. They are the nearest training-data point of any class that form the two parallel lines to the hyperplane, being the distance between these lines, the greatest possible [13].

For this project, the MRI images available will allow the analysis of their voxels in order to obtain the individual information relative to each of the subjects. What is intended is to

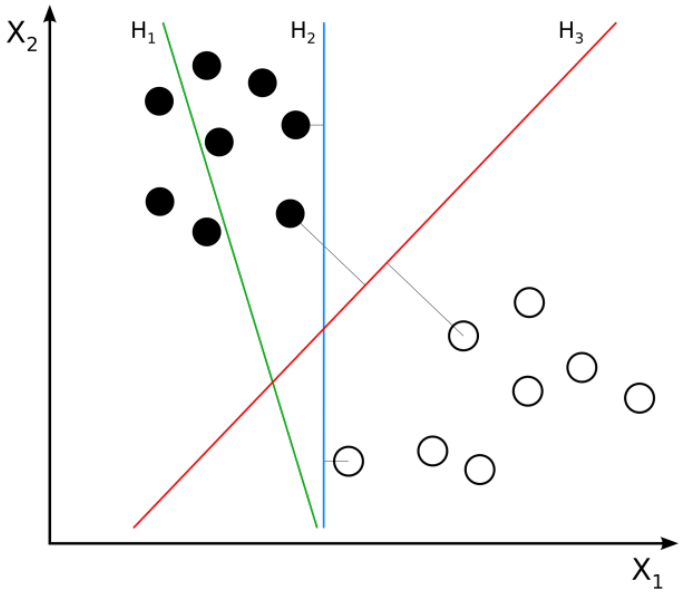


Figure 2.1: Infinite possible hyperplanes (H_1, H_2, H_3) in the classification. The hyperplane that gets the maximum distance between different input data is chosen, H_3 in this case. Image extracted from Wikipedia.

get an order and a distinction between those features, voxels, that represent a determinant contribution to be able to decide if a patient suffer Alzheimer or not. By means of the SVM, it is pursued to find the optimum thresholds of decision without reaching an overfitting that will provoke that the model could not be generalize for other data.

Linearly Separable Case

The following derivation for the linearly separable case is valid for a binary classification problem, in our case Alzheimer patients (AD) versus Control subjects (C); all the obtained conclusions in this binary case would be extrapolated to a Multiclass SVM, generating as many classifiers as classes available.

Given the input data set $\mathbf{x} \in R^n$ belonging to the different classes $y \in \{1, -1\}$, our objective is to obtain the values of the parameters of the hyperplane: $\{\mathbf{w}, \mathrm{b}\}$, which originate the biggest margin to the hyperplane, as it was previously mentioned.

The best hyperplane, in the accuracy sense, for the classification of data will be the one in which the margin of separation between the neighboring vectors of both classes, AD & C, is maximum. For the treated case, that hyperplane is H_0, denominated as *hyperplane of optimum separation.*

$$\mathbf{w} \cdot \mathbf{x_i} = b \lesseqgtr y_i \quad / \quad \begin{cases} y_i = 1 & \forall x_i \in C_i = AD \\ y_i = -1 & \forall x_i \in C_i = C \end{cases} \Rightarrow$$
$$\Rightarrow \exists(\mathbf{w}^*, b^*) \quad / \quad \mathbf{w}^*\mathbf{x} + b^* = 0 \quad Optimum\, Hyperplane \tag{2.2}$$

H_0 is obtained by reaching the maximum margin between input vectors of the two different classes. This margin is limited by two parallel hyperplanes, AD & C, which contain at

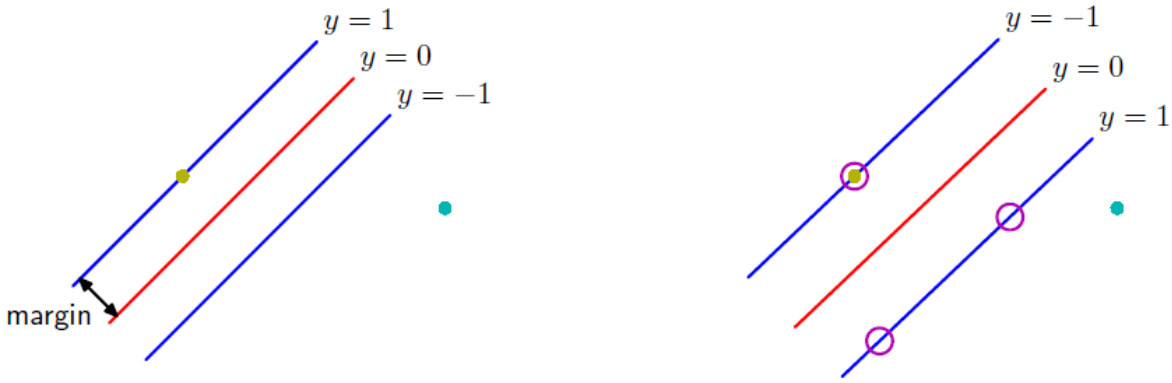


Figure 2.2: Left figure shows the margin as the perpendicular distance between the decision threshold and the closest point out of the set. Right figures represents how maximizing this margin leads to a particular choice of decision boundary. The points forming this boundary, indicated by circles, are known as support vectors. Extracted from C. M. Bishop (2006). Pattern Recognition and Machine Learning.

least one input vector of each class. These vectors, where one places the hyperplanes that limit the margin are called Support Vectors.

$$\begin{aligned} H_{AD} &: \mathbf{w} \cdot \mathbf{x} - b = 1; \\ H_C &: \mathbf{w} \cdot \mathbf{x} - b = -1; \end{aligned} \tag{2.3}$$

From the equation (2.3), the margin is defined as the distance between H_{AD} and H_C as,

$$d = \frac{2}{|\mathbf{w}|} \tag{2.4}$$

It is noticed that minimizing the norm of $\mathbf{w}$,$|\mathbf{w}|$, is equivalent to maximizing the margin between hyperplanes H_{AD} and H_C. Besides, it must be careful of not finding input vectors in the zone between both hyperplanes,

$$y_i(\mathbf{w}\mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n \tag{2.5}$$

This problem of optimization can be represented through,

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{||\mathbf{w}||^2}{2} \\ \text{s.t} \quad & y_i(\langle \mathbf{w} \cdot \mathbf{x_i} \rangle) + b) \geq 1 \end{aligned} \tag{2.6}$$

and illustrated by figure 2.3.

In mathematical programming, a problem such as (2.2) is called a convex quadratic problem. Many robust algorithms exist for solving the quadratic problems. Since the quadratic problem is convex, any local minimum found is always a global minimum [14].

For the resolution of this optimization problem it will be required the use of the Lagrange multipliers by means of the primal Lagrangian function,

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum_{i=1}^{l} \alpha_i [y_i(\langle \mathbf{w} \cdot \mathbf{x_i} \rangle + b) - 1] \tag{2.7}$$

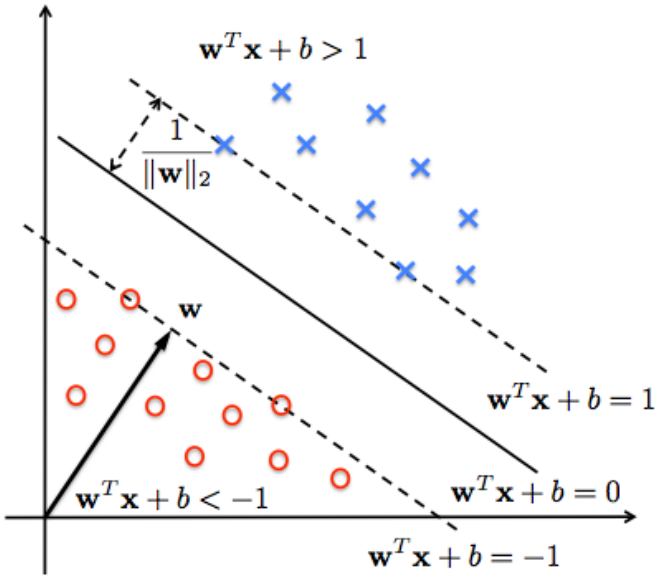


Figure 2.3: X symbol represents in our case the AD(+1) class and circles class Control(-1), limited by the boundary established by equations (2.6). Image extracted from Wikipedia.

Being $\alpha_i \geq 0$ the Lagrange multipliers. The problem will be solved if the input data vectors are linearly separable, ensuring that the minimum found is a global minimum, being the partial derivatives of Lagrange function equal to 0.

$$\begin{aligned} \frac{\partial L(\mathbf{w},b,\alpha)}{\partial b} &= \sum_{i=1}^{l} y_i\alpha_i = 0 \\ \frac{\partial L(\mathbf{w},b,\alpha)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{l} y_i\alpha_i\mathbf{x_i} = 0 \end{aligned} \tag{2.8}$$

This optimization scenario can be reformulated substituting the results obtained with the partial derivatives (2.8) into the primal problem (2.7) getting a dual formulation,

$$L(\mathbf{w},b,\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \mathbf{x_i}\cdot\mathbf{x_j}\rangle \tag{2.9}$$

Now we have the dual problem (2.9) corresponding to the the primal problem (2.7),

$$\begin{aligned} \text{maximizes} \quad & \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \mathbf{x_i}\cdot\mathbf{x_j}\rangle \\ \text{s.t} \quad & \sum_{i=1}^{l} y_i\alpha_i = 0,\ \alpha_i \geq 0,\ i = 1,...,l. \end{aligned} \tag{2.10}$$

The primal problem (2.7) and its corresponding dual problem (2.9) reach the same normal vector to the hyperplane,

$$\sum_{i=1}^{l} y_i\alpha_i\mathbf{x_i} + b \tag{2.11}$$

Linearly Inseparable Case: Soft Margin SVMs

In the previous section it is considered that in the region between the hyperplanes H_{AD} & H_C there are no input vectors (see figures 2.2 and 2.3) and none of them are misclassified. However, a perfect separation is not always possible, and in case of being perfect, that model can misclassify new data, producing overfitting [15].

To obtain flexibility, the SVM create an alternative to the maximum margin, the Soft margin, which, facing the situation of not obtaining an optimum hyperplane H_0 (2.2) of separation, will select a hyperplane that allows misclassification errors and, at the same time, maximizing the distance between the data that are correctly classified, shown in figure 2.4. To carry it out, it is needed to reduce each individual input data so a penalty variable must be introduced, the slack variable ξ and a constant parameter C, which will control the balance between training errors and overfitting [22].

$$y_i(w^T \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, ..., N. \tag{2.12}$$

where,

$$\xi_i \geq 0 \quad \forall i \quad \Rightarrow \quad \begin{cases} 0 \leq \xi_i \leq 1 \rightarrow \textbf{correct classification} \\ \xi_i > 1 \rightarrow \textbf{incorrect classification} \end{cases} \tag{2.13}$$

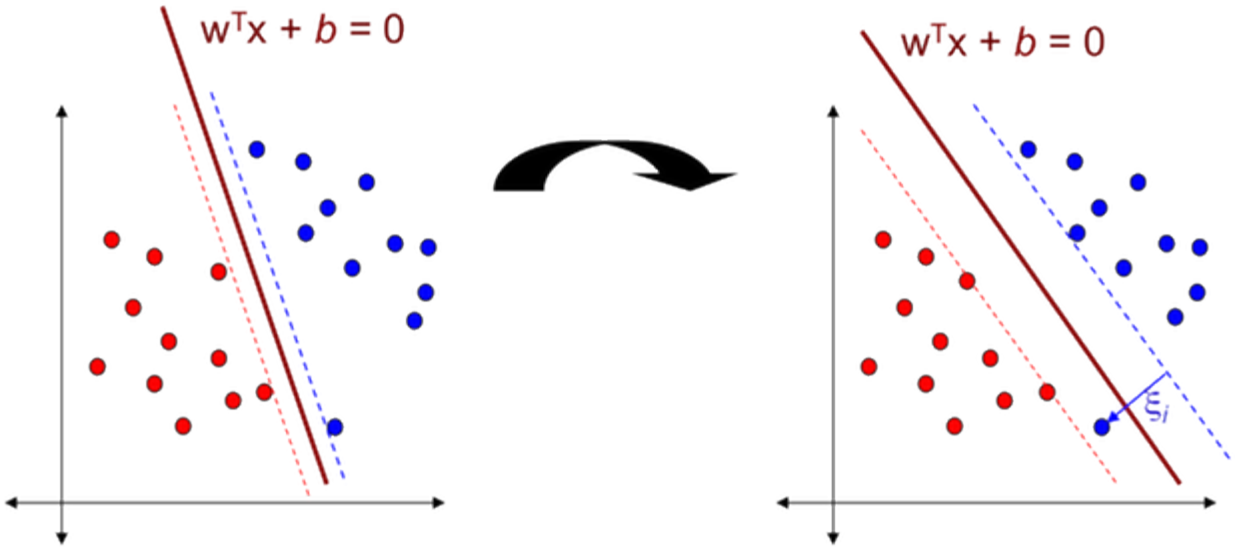


Figure 2.4: The classifier in the left is an optimum classifier, while the one in the right has become a linearly inseparable case with samples inside the margin area. Image extracted from http://courses.cs.ut.ee/.

We still looking for the maximum margin between hyperplanes and the minimum error, but with a modification with respect to the linear separable case (2.6), since now there is a dependency on the slack variable ξ, with that aim we must,

$$\begin{aligned} &\text{minimize} \quad \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{l}\xi_i \\ &\text{considering} \quad y_i(\mathbf{w}\cdot\mathbf{x_i} - b) \geq 1,\ 1 \leq i \leq n \end{aligned} \tag{2.14}$$

The primal Lagrangian function for the Linearly inseparable case is similar to the separable one (2.7),

$$L(\mathbf{w}, b, \xi, \alpha, \mathbf{r}) = \frac{1}{2}\langle \mathbf{w}\cdot\mathbf{w}\rangle + C\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i[y_i(\langle \mathbf{w}\cdot\mathbf{x_i}\rangle + b) - 1 + \xi_i] - \sum_{i=1}^{l}\mathbf{r_i}\xi i \tag{2.15}$$

Where $\alpha_i \geq 0$and $\mathbf{r_i} \geq 0$. Performing the partial derivatives of the Primal Lagrangian (2.15) and equalizing to 0, the dual formulation problem is reached,

$$\begin{aligned}
\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \mathbf{r})}{\partial b} &= \sum_{i=1}^{l} y_i \alpha_i = 0, \quad \Rightarrow \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \\
\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \mathbf{r})}{\partial b} &= C - \alpha_i - \mathbf{r_i} = 0, \quad \Rightarrow \quad C = \alpha_i + \mathbf{r_i} \\
\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \mathbf{r})}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{l} y_i \alpha_i \mathbf{x_i} = 0, \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x_i}
\end{aligned} \tag{2.16}$$

Substituting the partial derivatives (2.16) into the primal problem (2.15), the dual problem is obtained,

$$L(\mathbf{w}, b, \xi, \alpha, \mathbf{r}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \mathbf{x_i} \cdot \mathbf{x_j} \rangle \tag{2.17}$$

Again, as in the linearly separable case, the dual problem corresponding to the primal one (2.15) is presented,

$$\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \mathbf{x_i} \cdot \mathbf{x_j} \rangle \\
\text{s.t} \quad & \sum_{i=1}^{l} y_i \alpha_i = 0,\ C \geq \alpha_i \geq 0,\ i, j = 1, ..., l.
\end{aligned} \tag{2.18}$$

The constant C limits the influence of any point by being the upper bound for the Lagrange multiplier, α.

In this project it has been working with a higher number of features than subjects, so it has a lineal solution without being necessary that some of the data needed is misclassified. As it was explained before, the C parameter penalize to the misclassified data, so for the experiments it has been chosen a high value of C, getting a solution that classify all the subjects successfully. Nevertheless, an experiment to show the solutions depending on a different value of C has been implemented and will be explained and shown in 4.

2.3 Feature Selection

In order to solve the classification problems, using machine learning methods, the issue of the large number of input variables presented in the data must be faced. Working with the MRI images, the number of variables is extremely high, because the number of voxels is higher than $500,000$ in the MRI image, quite larger than the number of subjects, 425.

This extremely high number of variables causes some complications: complex models, which are hard to visualize and to understand by researchers, high measurement and storage requirements are needed, increment of training and utilization times and presence of redundant and/or irrelevant features which add noise rather than provide useful information for the classification [41]. All these drawbacks can be avoided by means of feature selection.

2.3.1 Introduction

A feature selection algorithm can be seen as the combination of search techniques for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets and find which of these subsets of features is the most relevant and informative. The three main categories of feature selection algorithms are: filters, wrappers and embedded methods [41].

- **Filter Methods**

  These methods select subsets of variables as a pre-processing step, regardless of the predictor. Applying some ranking over features, denoting how useful each feature is likely to be for classification, suppressing the least interesting variables. Once this ranking has been computed, a feature set composing of the best N features is created. However, filter methods tend to select redundant variables because they do not consider the relationships between variables. Therefore, they are mainly used as a pre-process method. These methods are particularly effective in computation time and robust to avoid overfitting [43].

- **Wrapper Methods**

  Unlike filter methods, wrapper methods allow to detect the possible interactions between variables, which is priceless, avoiding not just redundant variables, but being capable of selecting a variable useless/redundant by itself, but useful together with others. As the filters methods, a search algorithm is used to search through the space of possible features, but with the difference that wrapper methods use a predictive model to score feature subsets, and each new subset is used to train a model, producing a ranking based on the classification error obtained with each subset.

  This search method, with such an exhaustive selection of features, becomes computationally intractable and can be produced overfitting of the training data since the method tends to pick those features that produce good results with a given training set, but not necessarily in a more general scenario. Quite opposite to filter methods, which instead of evaluating against a model, a simple filter is evaluated, giving a lower prediction performance than a wrapper. In some occasions filters are used as a preprocessing step for wrapper methods.

  Wrapper drawbacks can be alleviated by means of forward selection and backwards elimination. Known as Greedy search strategies, the first one incorporate variables

progressively into larger and larger subsets, while the second one starts with the full set of variables and progressively eliminates the least valuable ones [41].

- **Embedded Methods**

  Proposed to reduce the classification of learning, the embedded methods are a combination of the advantages of the previous methods, but performing variable selection in the training process, so it needs a preliminary knowledge of what a good selection is. It follows an efficient process by implementing an optimization of a two-part objective function with "a goodness-of-fit term and a penalty for a large number of variables" [41]. The embedded techniques performed, in terms of computational complexity, between filter and wrappers methods, being less prone to overfitting than this last one.

In this project, we have chosen two feature selection algorithms to compare: the *t*-test, a filter method, and a Recursive Feature Elimination (SVM-RFE) algorithm, which is consider an embedded or wrapper method depending on the literature. Both are going to be fully described in the following sections. These two methods appear as the most common approaches to automatic feature selection in the MRI literature.

2.3.2 Univariate tests: *t*-test

A *t*-test is "any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution if the null hypothesis is supported" [49].

This univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values, describing the pattern of response of each variable on its own, individually [49].

Following the interest in identifying the features, which are better indicators in classifying patients of Alzheimer versus Control subjects, neuromarkers, the statistical hypothesis test, *t*-test, will be used. Specifically, a two-sample *t*-test. To justify the used of the *t*-test2 instead of *t*-test, the difference between both statistical hypothesis tests must be explained. This difference is mainly defined by the related sample sets employed, being for the*t*-test2 a scenario with independent or unpaired samples, while for the *t*-test paired or dependent samples are involved.

- **Independent or unpaired samples**

  "The independent samples *t*-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared". It is commonly applied when the statistical units underlying of the two samples being compared are non-overlapping [30]. So, a two-sample location test of the null hypothesis such that the means of two populations are equal (null hypothesis).

  From the point of view of the scenario of this project, it is wanted to identify if a patient suffers of Alzheimer or is a Control subject, for that, around 400 subjects are enrolled, 200 subjects are assigned to the patients group and 200 subjects to the control group. In this case, there are two sets of independent samples and would use the unpaired form of the two-sample *t*-test.

- **Paired samples**

  In the case of the paired samples t-tests, commonly consist of a sample of matched pairs of similar units, or one group of units that has been tested twice, so it can be also known as a "repeated measures" t-test.

  An example of the repeated measures t-test in the scenario of classifying Alzheimer's patients, would be where subjects are tested prior to a treatment, in case of being one available and successful, and the same subjects are tested again after the treatment. By comparing the same patient's numbers before and after treatment, it is used each patient as their own control. That way of correcting the rejection of the null hypothesis is more possible to take place, with statistical power increasing simply because the random between-patient variation has now been eliminated. But due to this increase of the statistical power, more tests are required, each subject having to be tested twice [26].

  It can be seen that from an unpaired sample that is repeatedly used, a paired samples t-test is obtained based on a "matched-pairs sample" results.

As it has been showed, a two-sample t-test is the statistical hypothesis test required since fits our model, being the practical process as follows.

In the first step, the data set is separated between patients and controls into different populations. For the function, is mandatory the same number of features in each subject, not the same number of subjects in each population. Then, the two-sample t-test is performed,

$$[H, p] = ttest2(AD, C); \tag{2.19}$$

With the hypothesis that the data from the two populations come from independent random samples from normal distributions with equal means and equal, but unknown, variances. This is known as the null hypothesis and is said to be accepted. The concept is graphically represented in figure 2.5.

The alternative hypothesis is that each population has a different mean, shown in figure 2.6. The function returns the result of the test in H. If the value obtained of H is 0, the null hypothesis is true, equal means. Otherwise, if H is equal to 1, corresponds to a small p-value, which casts doubt on the validity of the null hypothesis, being rejected at the 5% significance level [2].

As it is desired to find between all the features, the ones which distinguish between both populations, the ranked list of features will be based on the p-value of the t-test, being the features with small p-values at the top for the characterization between Alzheimer patients and Control subjects.

2.3.3 Multivariate tests: SVM-RFE

Continuing with the aim of finding which voxels of the MRI images are relevant to distinguish between Alzheimer patients and Control subjects, the Recursive Feature Elimination

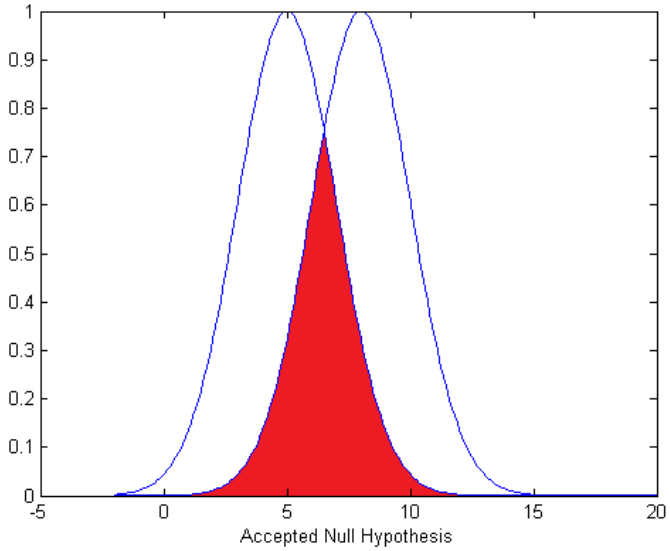


Figure 2.5: The accepted null hypothesis, consequence of two distribution with equal or quite closed means, where is hardly possible to identify samples from two different populations.

is implemented as an iterative feature selection algorithm developed specifically for Support Vectors Machines [40]. Specifically developed for the SVM since the RFE algorithm selects features depending on the classification margin supplied by the classifier of the SVM, there is not a separation between the learning process and the feature selection one.

The SVM-RFE represents a multivariate testing technique for testing a hypothesis in which multiple variables are modified. The goal of multivariate testing is to determine which combination of variations performs the best out of all of the possible combinations [44].

The iterative procedure followed by the RFE is basically based in three steps:

1. Training of the linear classifier optimizing the weights $\mathbf{w_i}$.
2. Establishing of the ranking criterion for all the features of each subject.
3. Removing of the feature corresponding with the lowest value of $|\mathbf{w_i}|$, since is the one which the variation in the classification margin is the smallest.

The feature set will be decreased in smaller subsets sequentially until the last feature is eliminated, being this removal process going parallel to the ranking feature one, and continued until no more variables are left.

The whole algorithm of the SVM RFE is an application of the RFE in the linear case, using the weight magnitude as ranking criterion [42]:

1. Given training instances $\mathbf{x_{all}} = \{x_1, ..., x_N\}$, and class labels $y = \{y_1, ..., y_N\}$, initialize the subset of features $\mathbf{s} = [1, 2, ..., n]$ and $\mathbf{r} = []$ an empty array.
2. Repeat (a)-(f) until $\mathbf{s}$ becomes an empty array, $\mathbf{s} = []$.

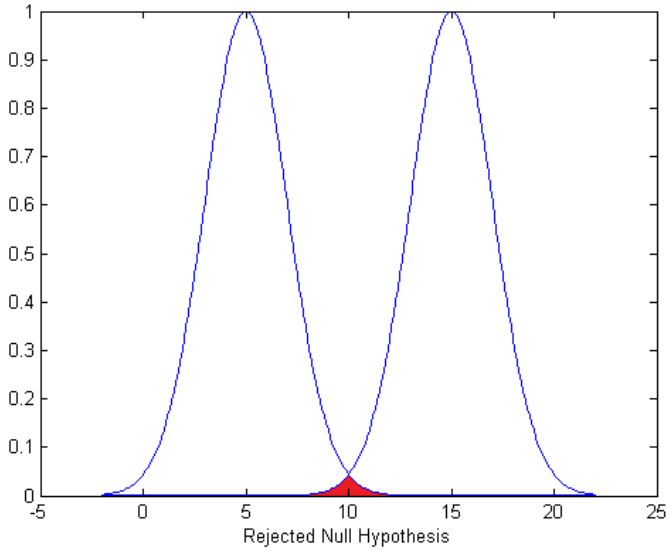


Figure 2.6: The rejected null hyphotesis. The null hypothesis can be rejected since the means of the two populations under study are different.

(a) Construct new training instances $\mathbf{x} = \mathbf{x_{all}}(:;\mathbf{s})$

(b) Train α=swmtrain ($\mathbf{x}$,y)

(c) Computing the weight vector of dimension l=length (s),

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x_i} \tag{2.20}$$

(d) Find the feature f with the smallest $\mathbf{w_i}; i = 1, 2, \ldots, |s|$ $f = \arg\min(|w|)$.

(e) Update $\mathbf{r}$,

$$\mathbf{r} = [\mathbf{s}(f), r] \tag{2.21}$$

(f) and eliminate the least relevant feature from $\mathbf{s}$,

$$\mathbf{s}(f) = [] \tag{2.22}$$

3. Obtaining as output the feature ranked list $\mathbf{r}$, being The last eliminated feature, the most relevant one.

The described procedure is an instance of backward feature elimination [40]. It has been explained following a one single feature elimination per iteration, but in order to reduce the computational cost and increase the speed of the whole process, more than one feature can be removed in each step.

As a difference with respect to the t-test (filter methods of feature selection), the selected features are not the ones which individually classify best the training data, since correlation and expression ratio methods are included in the RFE.

For the scenario of this project, there will be a first iteration used to select the best $100,000$ variables out of the more than $500,000$, eliminating the rest, which corresponds with the lowest weight value. Then, during the training, every iteration will rank and eliminate variables in steps of $5,000$ variables. Despite the fact that performing in this way can reduce the accuracy of the classification, this method will provide subsets of samples that will finish faster than eliminating individually.

2.4 Performance Evaluation

As important as the training methods, is the performance evaluation. One usual performance measurement for classification problems is the 0/1 loss function [47], i.e., percentage of misclassified samples. However, it is very important to remark that our objective is not to minimize this cost over the training data but to evaluate how good our approach classifies test samples, i.e. how well our method generalizes to new data. Consequently we have to use some technique that allows us to approximate that performance from the labeled data we have.

2.4.1 Introduction

There are two accuracy estimation methods which are the most common ones, **Conventional Validation** and **Cross-Validation**.

In the first one, also known as *holdout*, the available instances of the dataset are portioned into two subsets, a training set and a test set, also known as *holdout* set. Commonly the distribution of the data between both is 2/3 of the data instances designated for the training set and 1/3 for the test set [47].

This practice makes an inefficient use of the data for several reasons. The data size is always finite, and in the majority of the cases, smaller than it would be desired, so not using a 1/3 of the dataset for training will be senseless and will increase the severity of this small sample problem even more. Besides, as it has been mentioned many times during this memory, the issue of overfitting would be highly probable. The cross-validation is able to perform overcoming the drawbacks of the *holdout* method [47].

2.4.2 Cross-Validation

The cross-validation, also known as rotation estimation [47][34][27], defines a dataset to test the validation model during the training phase in order to estimate the accuracy of the classifier giving a view of how the model will generalize to an independent dataset.

This technique will evaluate the statistical analysis results and will guarantee the independency between training and test data. As the conventional validation, the two subsets of data are used to perform the cross-validation analysis, which is the arithmetic mean of the obtained evaluation measures over different partitions, but with the difference that many rounds, changing the data belonging to the partitions, are executed, getting a reduction of the variance and being the validation results the average over the iterations. This behavior enables a more accurate estimation of the model prediction performance [58].

The cross-validation can be classified in two types, non-exhaustive and exhaustive cross-validation.

- **Non-exhaustive cross-validation**

  These cross-validation methods do not compute all ways of splitting the original data sample. The *k-fold* cross-validation performs a cross-validation of k iterations where data sample is divided in k subsets. One of these subsets is used as a validation

subset and the others *k-1* as training sets, changing the validation subset in each iteration, so that every subset is a validation set once. At the end of the process the arithmetic mean of the results is performed to reach a unique result. The election of the value of k depends on the samples size, being the most common the one of 10 iterations, *10-fold* cross-validation [46][28] and the simplest one, the *2-fold* cross-validation, where data samples are divided into two equal size subsets, after being shuffle. Then, 2 iterations are performed, rotating the training and validation subset in each iteration.

Exists another cross-validation method belonging to the non-exhaustive type, the *repeated random sub-sampling* validation, consisting in dividing randomly the dataset into training and validation data. Having these two splits, the model is fit to the training data, and predictive accuracy is assessed using the validation data. The unique result of the validation is got by the arithmetic mean of the results.

One advantage with respect to the *k-fold* cross-validation is that the division of the dataset does not depend on the number of iterations. But there is also a drawback, since some samples of the dataset can be not evaluated and others can be evaluated more than once, causing an overlapping of the training and validation subsets [50]. This method is also known as Monte Carlo cross-validation [54], due to the generation of Monte Carlo samples of the data set. The result of this analysis will differ when repeated, because of random splits.

- **Exhaustive cross-validation**

  Unlike the non-exhaustive, the exhaustive cross-validation methods compute all ways of splitting into a training and a validation set. The methods that represent this type are the *leave-p-out* and the *leave-one-out* ones.

  The *leave-p-out* cross-validation uses p observations as the validation set and the remaining observations, *N-p*, as the training set. Comparing with methods seen previously, can be appreciated that the *k-fold* cross-validation is an approximation of the **LpO CV**. Also, when the number of random splits goes to infinity, the *repeated random sub-sampling* validation become arbitrary close to the *leave-p-out* cross-validation.

  Assigning a value of 1 to the number of observations, p, the *leave-p-out* cross-validation becomes a *leave-one-out* cross-validation. This method splits the data assigning one sample to the validation set, and the rest, *N-1*, to the training set. There will be N iterations, rotating the sample of the validation set in each one. This is the most accurate method with the lower error but with the biggest computational cost in comparison with the rest of methods seen due to the high number of iterations performed, as many as samples we have, N [1].

2.4.3 Leave-one-out

In our project the *leave-one-out* cross-validation (**LOOCV**) will be the performance evaluation employed. The reason is that it maximizes the amount of available data for training getting a better accuracy of the SVM classifier, despite the computational cost.

In practice, as mentioned above, the way in which we will use the **LOOCV** is taking the whole amount of subjects we have available, 425, and take one different sample in each iteration as validation sample and the rest for training.

The following code represents the process:

1. For loop that will implements as many iterations as subjects are available,

$$\text{for } ii = 1 : suj \tag{2.23}$$

2. Assign all the sample dataset to an array,

$$pos\_t\_xsel = (1 : suj); \tag{2.24}$$

3. From this array, the validation sample is subtracted and kept in a different one, splitting the dataset in a validation subset of one sample and a training subset of $N - 1$ (425-1),

$$pos\_v\_xsel = pos\_t\_xsel(ii); pos\_t\_xsel(pos\_v\_xsel) = []; \tag{2.25}$$

4. Now the SVM linear classifier proceeds to optimize the model parameters to make the model fit the training data the best as possible, then the generalization of the model will be done by the independent samples of the validation subset, just one sample in our case.

Chapter 3

Data acquisition and preprocessing

The input data space employed in this project has been acquired from the ADNI database [5], introduced in section 1.3.

Out of all the available clinical data and subjects belonging to all the phases of the Alzheimer evolution process, we have chosen the MRI images of the Alzheimer and Control subjects. We decided to select these two classes for the study instead of the also available EMCI[1] and LMCI[2] to limit the scope of this project and for considering them more clearly differentiable in the classification, anyway they will be suggested for future lines of research in chapter 5.

In this part of the project we have focused on selecting the highest number of samples with the same imaging characteristics, acquiring a similar number of AD and Control, and following the preprocessed methods needed to optimally prepare the data samples for the training model in order to obtain the highest accuracy.

3.1 Structural Magnetic Resonance Imaging

Since our objective is to find the biomarkers of the brain regions that are most relevant to Alzheimer, we define the structural MRI (sMRI), since it provides anatomical reference for visualization of activation patterns and regions of interest in the brain to extract functional signal information from Alzheimer and Control subjects.

The Magnetic resonance imaging (MRI), is a medical imaging technique used in radiology to image the anatomy and the physiological processes of the body in both health and disease. MRI scanners use strong magnetic fields, radio waves, and field gradients to form images of the body [8]. The structural MRI provides information that is used to describe the shape, size and integrity of different tissues in the brain.

The main reason of using the structural MRI images is that are highly sensitive for detecting relevant neurodegenerative changes in Alzheimer disease and are used as an outcome measure for clinical trials [8].

MRI scans are considered to be a safe procedure, providing you do not have any implants

[1] Early Mild Cognitive Impairment
[2] Late Mild Cognitive Impairment

or metallic objects on you that must not go in the scanner. As a big advantage, it does not use radiation and they are therefore suitable for use in children and pregnant women.

It is a quite recent technology, since it was not possible to get the first clinically useful image of a patient's internal tissues until August 1980, identifying a primary tumor in the patient's chest, an abnormal liver, and secondary cancer in his bones [9].

The process begins with the patient/volunteer being positioned inside a MRI scanner, similar to the one in figure 3.1. The scanner forms a strong magnetic field around the area to be imaged, the brain in our case. The hydrogen atoms (protons) contained in the water molecules of the brain tissues are used to create a signal that is preprocessed to form an image of the brain.

Figure 3.1: MRI scanner.

The energy from an oscillating magnetic field is temporarily applied to the patient at the appropriate resonance frequency. The excited hydrogen atoms emit a radio frequency signal which is measured by a receiving coil. The radio signal can be made to encode position information by varying the main magnetic field using gradient coils. As these coils are rapidly switched on and off they create the characteristic repetitive noise of an MRI scan. The contrast between different tissues is determined by the rate at which excited atoms return to the equilibrium state. The elements which formed the MRI scanner are indicated in figure 3.2.

Normally not used, but sometimes, for more specific types of imaging, contrast agents are needed and for MRI are based on chelates of gadolinium. In general, these agents have proved safer than the iodinated contrast agents used in X-ray radiography, with a rare probability (0.03–0.1%) of generating an Anaphylactoid reactions [7]. These contrast agents are introduced intravenously, orally or intra-articularly into the patient [51]. MRI requires a magnetic field that is both strong and uniform. The field strength of the magnet is measured in teslas – and the majority of systems operate at 1.5T [48].

Image contrast may be weighted emphasizing different aspects of normal and abnormal brain tissue to demonstrate different anatomical structures or pathologies. By modifying the sequence parameters of repetition time (TR) and echo time (TE), for example, anatomical images can emphasize contrast between gray and white matter, T1-weighted or between brain tissue and cerebrospinal fluid, T2-weighted.

The T1-weighted anatomical image 3.3, with short TR and short TE, provides a good contrast between gray matter (dark gray) and white matter (lighter gray) tissues, while CSF[3]

[3]Cerebrospinal Fluid

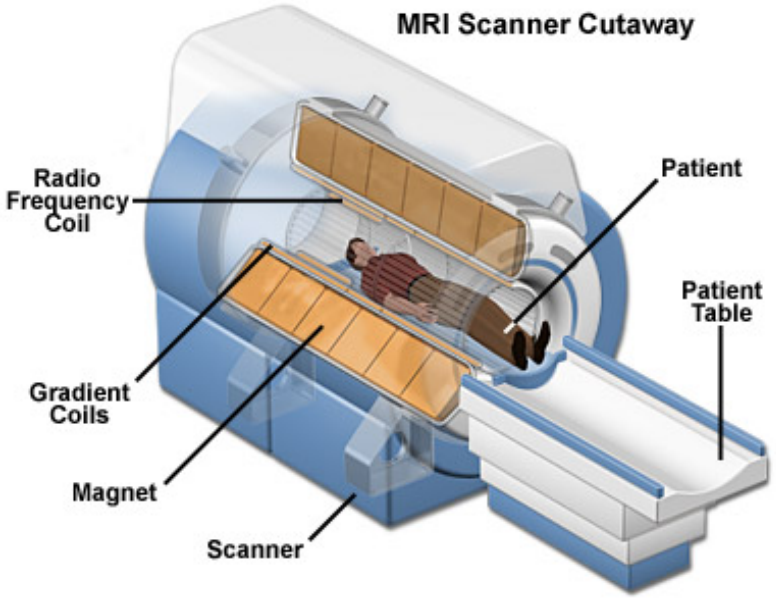


Figure 3.2: Parts of a MRI scanner.

is void of signal (black). Pathological processes, such as demyelination or inflammation, often increase water content in tissues, which decreases the signal on T1; white matter disease often shows up as darker areas in the lighter gray-colored white matter. Due to a better measure of water content, T2-weighted images are more sensitive to subtle white matter alterations [4].

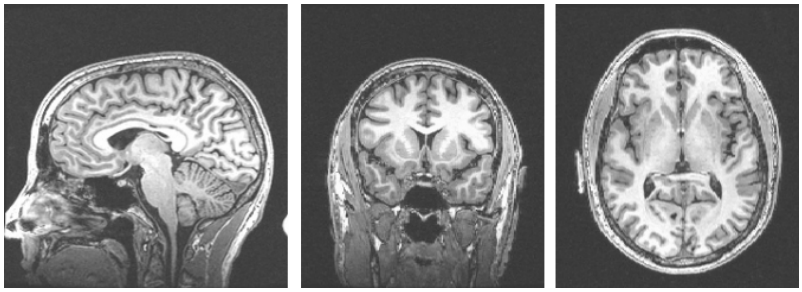


Figure 3.3: Three T1-weighted anatomical image from different sides, providing good contrast between gray matter (dark gray) and white matter (lighter gray). CSF is difficult perceptible (black).

The T2-weighted anatomical image 3.4, with long TR and long TE, provides good contrast between CSF (bright) and brain tissue (dark). Some T2 sequences demonstrate additional contrast between gray matter (lighter gray) and white matter (darker gray). Pathological processes, such as demyelination or inflammation, often increase water content in tissues, which increases signal on T2; white matter disease often shows up as brighter areas, which makes subtle changes easier to detect [4].

The result of the scanning process is a 3-dimensional image made up by voxels. A voxel (combination of the words “volume" and “pixel".) in a 3-D image is analogous to a pixel in a 2-D image. It is a small cube of brain tissue that represents around a million of brain cells [63]. Figure 3.5 is a 3-D brain image represented by voxels and how it looks after a smoothing process.

The value of each of these voxels corresponds to the density of the brain tissue in that point of the space. For example, in a T1-weighted MRI the gray matter (dark gray)

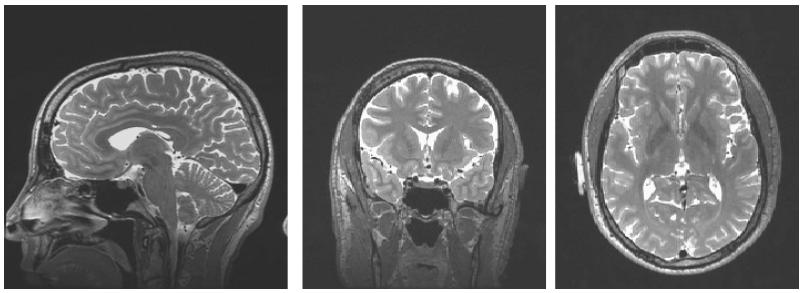


Figure 3.4: Three T2-weighted anatomical image from different sides, providing good contrast between CSF (bright) and brain tissue (dark).

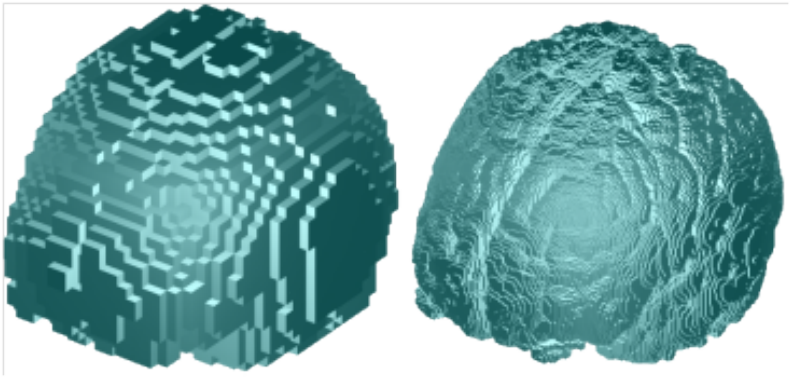


Figure 3.5: Left figure shows a brain image made by voxel, and right figure, shows its new aspect after a smoothing process. Image extracted from http://3dfd.ujaen.es/.

and white matter (lighter gray) tissues, while CSF (black), the low values of the voxels corresponds to the GM and WM wanted. We can analyze structural variations in the tissue by observing density variations from voxel to voxel. Figure 3.6 is a good example of a MRI brain scanner where voxels with the same density are colored.

3.2 Samples Selection

The neuroimages we are going to use from the ADNI database are basically the T1-weighted anatomical MRI brain scans in 3-D with a field strength of 1.5 Tesla. The reason of using the T1-weighted MRI is that, as mentioned in 3.1, "provides a good contrast between gray matter (dark gray) and white matter (lighter gray)" because are the regions in the brain where the loss and deterioration of the superior brain functions consequence of Alzheimer's disease are easier to appreciate [31].

The MRI images with 1.5T strength are chosen despite the fact the ones with 3 Tesla are available. The 3T MRI has a better SNR[4] and a higher SAR[5] [59][33]. Even though the better characteristics of 3T MRI, the number of 1.5T MRI available in the database is much higher, being this circumstance determinant for choosing them in our project. The samples with the wanted characteristics belong mostly to the ADNI1 phase, presented in figure 1.2.

All of these MRI images has been downloaded in a NIfTI[6] format from the ADNI database,

[4]Signal-to-Noise Ratio
[5]Specific Absortion Rate
[6]Neuroimaging Informatics Technology Initiative

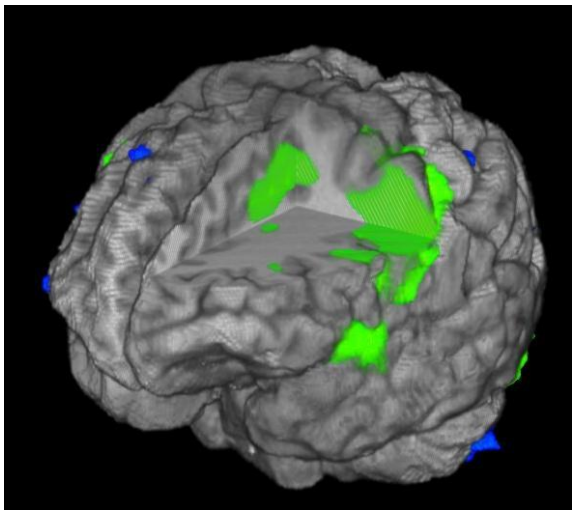


Figure 3.6: Voxels which correspond to the same density are colored in this MRI brain scan. Image extracted from http://www3.gehealthcare.com/.

by the advance search, shown in figures 3.7 and 3.8, in the official web [5].

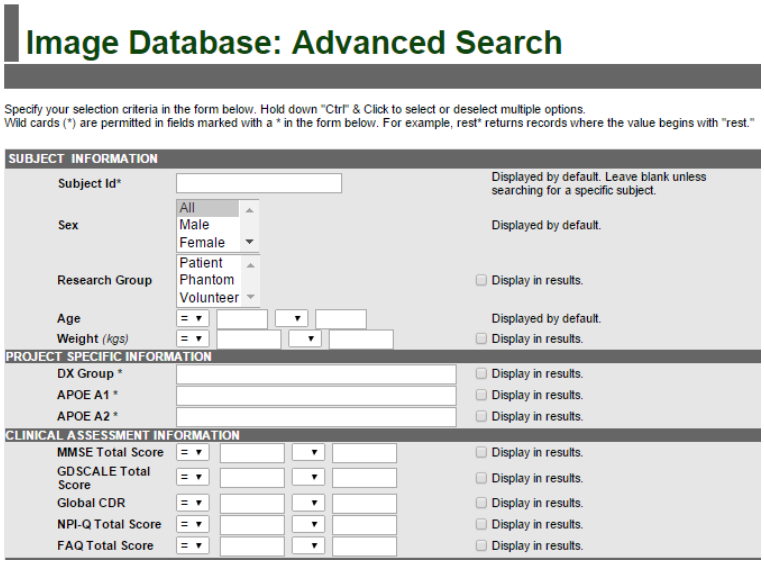


Figure 3.7: GUI of the Advance Search. "DX Group" refers to the possible labels of a subject: Control, EMCI, LMCI, AD. Extracted from ADNI webpage.

Once the download has finished, we have obtained 195 AD samples and 230 Control samples, being a total set of 425 samples. Each of these samples is a directory with several .nii images taken in different time periods and a XML file (ref appendix) with all the characteristics as subject ID, subject physical characteristics, cognitive tests scores (CDR[7], GDS, MMSE[8],...), technical image characteristics and the label with the corresponding phase of the process 1.2.

For this project we have taken as a reference the paper "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images" [19]. The reason is because the statistical analysis they have carried out are similar to what we have planned to do and they have used the Alzheimer's samples from the database as well. Comparing their project and ours, we have

[7] 0.5-very mind; 1-mild; 2; moderate; 3-severe
[8] Mini-Mental State Examination (Sccores 0-30)

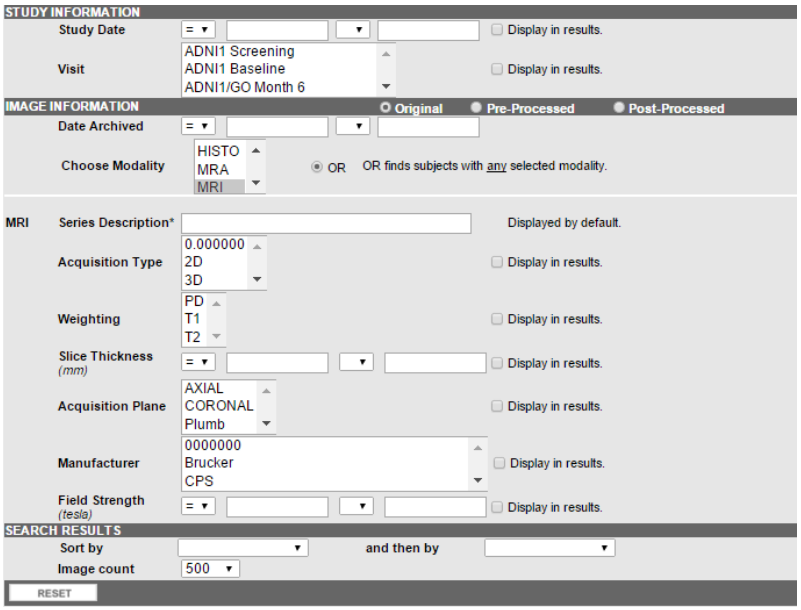


Figure 3.8: GUI of the Advance Search. "Weighting" we want a high T1-weighted anatomical image; "Field Strength (*Tesla*)" is going to be the 1.5T . Extracted from ADNI webpage.

been capable of getting a higher number of valid samples for the study. It is worthwhile mentioning that it has been a great guide for knowing what kind of results we could expect from our trials. This is why we decided to choose the earliest MRI in time between all the available ones for each subject, based in what they had decided in their article, so that we had a common start point. To filter this first MRI image taken from the others in each of the 425 subjects, a Java script was implemented.

3.3 Visualization & Preprocessing

MATLAB (**MAT**rix **LAB**oratory) is the mathematic software tool that we have employed to develop each of the experiment scripts and visualize images and figures. It has been chosen due to its capability for matrix manipulation with a high precision of calculation, reliable representation of data, functions and algorithms implementation and its big number of available toolboxes. It has its own programming language, M [2][36].

Figure 3.9 shows the visualization of the NIfTI (.nii) images downloaded from the database by means of the NIfTI toolbox of MATLAB implemented by Jimmy Shen [3], which was installed in our framework.

These images have to be preprocessed to transform them into a new space of variables in order to solve in an easier way the pattern recognition problem. The preprocessing process has been implemented with the Statistical Parametric Mapping (SPM12). The SPM12 is an academic software toolkit for used for the analysis of functional imaging data [56].

The preprocessed process consists on: **Re-orientation**, **Segmentation** and **Normalization**.

- **Re-orientation**

  The images of the 425 subjects were manually re-oriented to set the origin of their coordinate system as close as possible to the anterior commissure (ac).

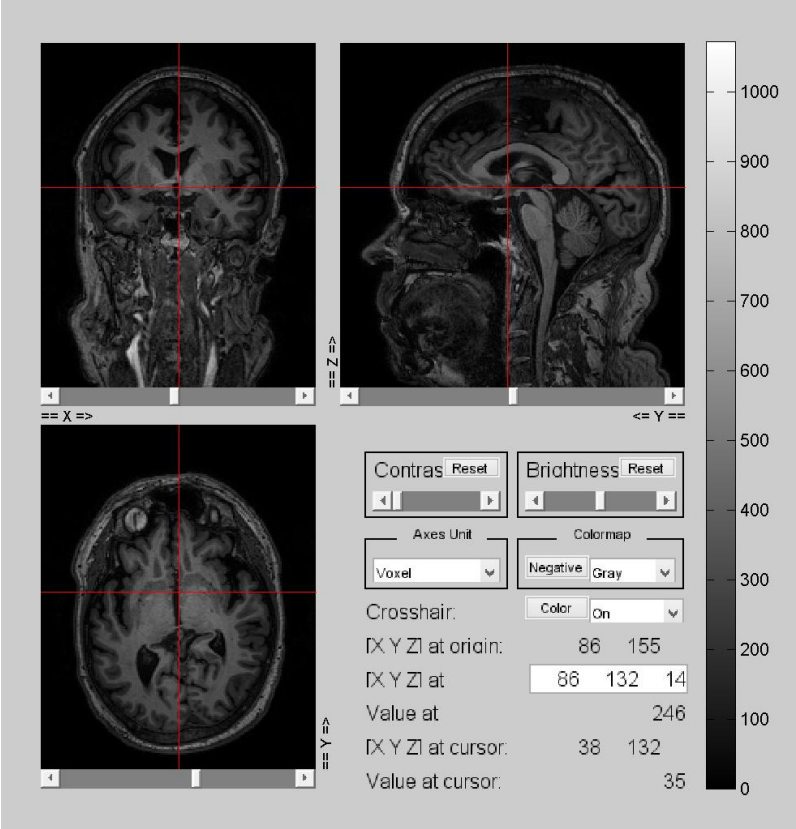


Figure 3.9: Visualization of a MRI image of a subject by NIfTI toolbox of MATLAB.

The **ac** is one of the two commissures existing in the brain, the other one is the posterior commissure (pc). Both are fiber tracts connecting the two hemispheres. The **pc** "connects the midbrain and diencephelon structures, lying just in front of and above the superior colliculli, below the pineal gland". It is not easy to see it on high detailed structural scan. The ac "connects the middle and inferior temporal gyri of the two hemispheres and runs across the middle just in front of the formix", being quite simple to distinguish in structural scans [61][29].

Then we set the origin of the scanner as *Talarich space* [61][29], setting the brain so that the anterior and posterior commissures are on a horizontal line, graphically showed in Figure 3.10.

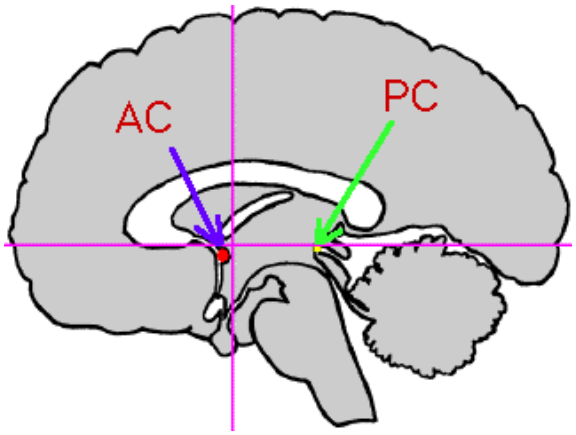


Figure 3.10: Origin set so that the ac and pc are on a horizontal line. Image extracted from "Statistical Parametric Mapping: The Analysis of Functional Brain Images".

MRI brain scan, in figure 3.11, is visualized with SPM12 GUI and easily set to the

origin by moving the axes and re-orienting them.

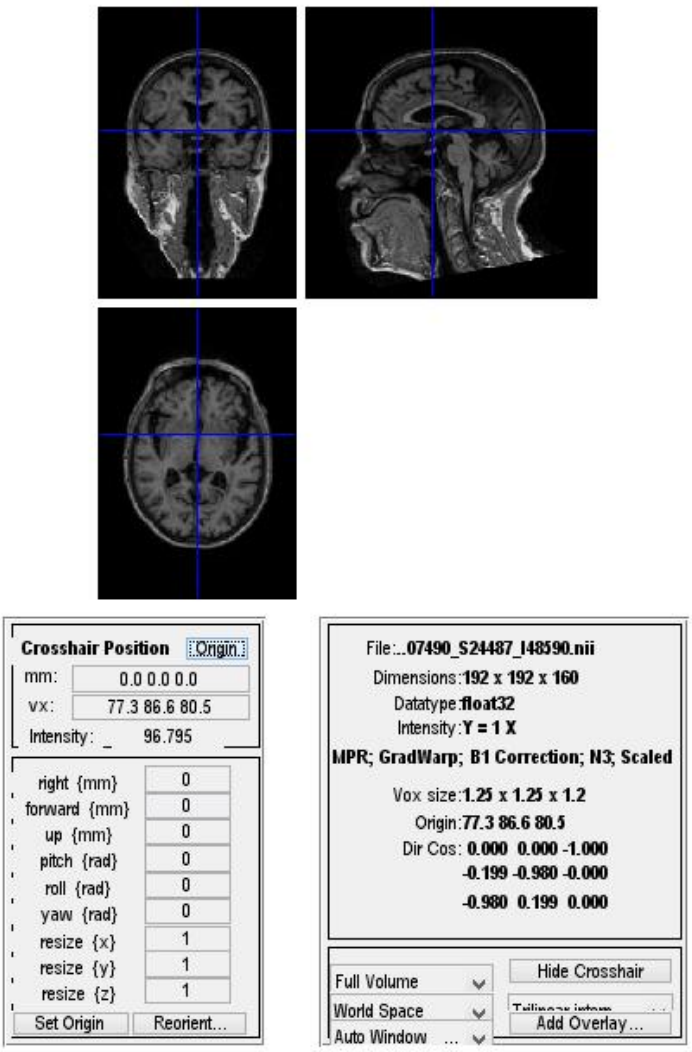


Figure 3.11: Origin set to the anterior commissure in the SPM12 GUI.

- **Segmentation**

  The reoriented images are then segmented, dividing the brain into the different types of tissue it is composed of, in our case are 5 segments tissue and they are showed in figure 3.12. The SPM is able to perform the segment by its "new segment" toolbox. We have followed the standard's settings, which also uses a 6 mm isotropic smoothing, which is standard in the literature.

  Segmentation provides a simplification of the image analysis and its interpretation. In our project we are interested in the grey matter (GM) segment because they are the most relevant to the detection of the Alzheimer's disease [32].

- **Normalization**

  After the Segmentation comes Normalization. It is important to highlight that, when working with MRI brain scans, the analysis carried out to each individual subject is done under the assumption that its voxels are placed in the same anatomical regions as in the rest of images. Obviously, no two brains are equal, so an standardization has to be performed. The normalization adjust the anatomy of each subject's brain to a common brain framework. Doing a brief explanation, an average of the GM

images of all subjects is implemented, generating an standardized brain template. This template is used to map all the brains to a common space. An example of this normalization is showed in the bottom right corner of the figure 3.12. This is possible thanks to *Dartel* toolbox [10].

At the end of the process, these masked, warped and modulated images, with $521,389$ voxels each, are then vectorized and become the input features for the feature selection and classification procedures developed in this project.

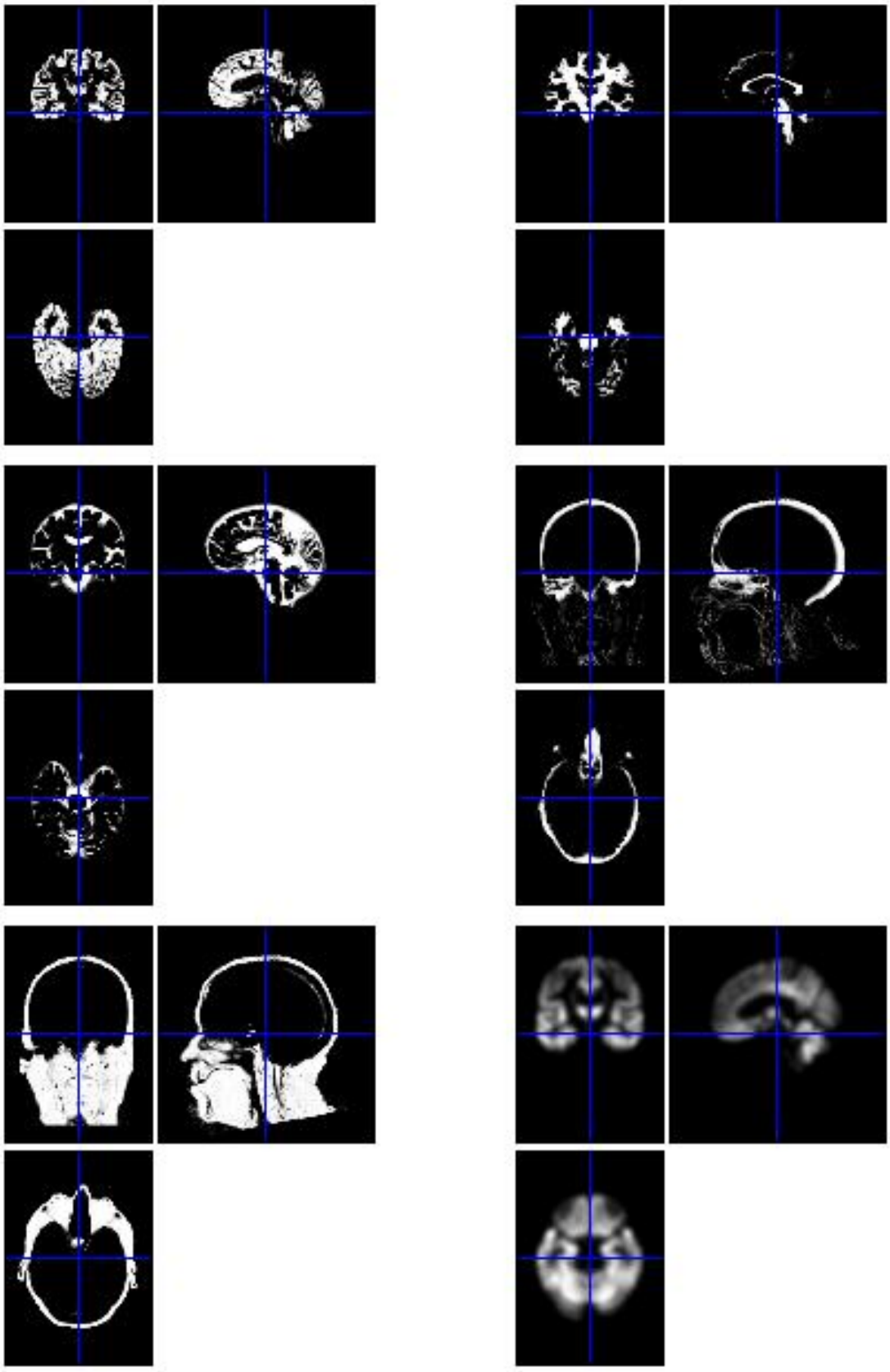


Figure 3.12: SPM12 GUI representation of the 5 different types of tissue in the segmentaion and the final MRI normalized: Grey Matter (top left corner), White Matter (top right corner), CSF (middle left), Cranium (middle right), Hair (bottom left corner), Normalized MRI ready for vectorization (bottom right corner).

Chapter 4

Analysis of MRI with machine learning methods

This chapter shows and explains the results obtained after applying the methods described in chapter 2 with the downloaded and preprocessed samples presented in chapter 3.

It is important to remark that all the algorithms described in chapter 3 has been firstly designed and executed with a Synthetic dataset, made up by ourselves, and then with ADNI dataset. The reason is explained in 4.1.2.

All the algorithms has followed a leave-one-out approach, which is the specific validation and testing strategy that was used and has been previously described in section 2.4, specifically in subsection 2.4.3.

In this chapter an analysis of the performance of each of the feature selection strategies seen in section 2.3 is performed to determine which is the most effective combination. We also present the evolution of the classification test error with respect to the number of selected features.

Finally, the most relevant features for the two implemented feature selection algorithms, are mapped to a brain template and rendered so that they may be visually analyzed showing the region in the brain more relevant for Alzheimer detection in a patient.

The algorithms were designed and tested on a HP Pavilion 15 NotebookPC laptop running Matlab 8.0.0.783, R2012b (The MathWorks Inc, Natick, Mass) operating on a Windows system. The Matlab SVM classification library that was used is libSVM (www.csie.ntu.edu.tw/ cjlin/libsvm). The data processing of a high grade of computation, was performed on uc3m machine, amaterasu.tsc.uc3m.es. The results were brought back to the HP Pavilion 15 NotebookPC laptop for analysis and visualization purposes. Finally MRI scans were processed and visualized using MRIcron (www.MRIcron.com).

4.1 Synthetic Experiment

4.1.1 Data description

The Synthetic data is a dataset employed to develop and test the algorithms that will be, afterwards, implemented using the prepared ADNI data collection. The set simulates the characteristics of the ADNI one but is easier to handle. The reason of using it is due to the high dimensionality, 521, 389 features after the vectorization, of the ADNI MRI images obtained from the preprocessing process explained in section 3.3. This high-dimensionality causes a high computational cost, meaning large times of execution of the scripts, the pc getting hung continuously, and much time needed to test every change in the code.

As a consequence, the performance would not succeed. Besides, for a better understanding of the algorithms, we decided to do it first in a simple data set like the Synthetics one, as we considered that it would help for a better concept understanding at the beginning of every new experiment and it provides a more clear interpretation before doing it in the data set of interest. In this chapter we will reference the experiments carried out with the synthetic dataset as "Synthetics experiments".

For the Synthetics experiments we a have a matrix X formed by 400 samples (rows) and 60 features (columns), X_{400x60}. Another matrix, a column one, is generated containing the labels for binary case, *+1* for the Alzheimer patients and *-1* for the Control subjects. This column matrix is Y_{1x60}.

Coming back to the X matrix, out of the 400 samples, 200 are simulating the Alzheimer patients, X_0, and the other 200 the Control subjects, X_1. Attending to the features of each sample, they are not all relevant, being just the first 20 useful for the classification and the other 40 are noise.

The feature's values are determined taking into account how the t-test and the SVM-RFE feature selection, from section 2.3, work.

For example, citing a paragraph corresponding to the t-test, *"As it is desired to find between all the features, the ones which distinguish between both populations, the ranked list of features will be based on the p-value of the t-test, being the features with small p-values at the top for the characterization between Alzheimer patients and Control subjects"*.

The relevant features (the first 20 columns) are generated as random Gaussian variables with a uniformly generated mean in the range (0,1), when the sample corresponds to a patient, and in (-1,0) when corresponds to a control. The variances of these Gaussian distributions are also uniformly distributed in the range(0,1). The noisy features (last 40) are simple zero-mean guassian noise of unit variance.

4.1.2 Performance analysis

After the introduction of the Synthetic data, we will analyze the results obtained for applying the t-test and SVM-RFE methods to them. First of all, remark that when training the SVM classifier, a C value, introduced in subsection 2.2.2, is needed to be chosen. The justification will be done in section 4.2 as it has been proved by ADNI dataset.

In the analysis of algorithms for *t*-test and SVM-RFE, we will pay special attention to the results obtained from a *rank order* vector which stores the ranking of all the features of each subject. This allows to see in which positions are located the best features for the classification of the samples. Besides, the accuracy % which measures the success of the classification.

Table 4.1 corresponds to the *rank order* vector of the *t*-test experiment. It is appreciable that the best 20 features for classification correspond to the first 20 features of the samples. This was expected and shows that the algorithm works fine, since as mentioned in subsection 4.1.2, the first 20 features are the only relevant ones, the rest are noise.

1	11	7	20	18	2	6	3	19	12	8	13	5	10	15	17	14	9	4	16

Table 4.1: The 20 features most relevant in Synthetic data by *t*test.

In table 4.2 of the SVM-RFE experiment, result is not as successful as in the *t*-test, since positions *42, 40, 37, 29, 36* corresponding to the *rank order* vector are ranked wrong for the reason explained in the previous paragraph.

13	2	5	20	9	18	10	3	**42**	1	14	8	**40**	19	**37**	17	15	**29**	11	**36**

Table 4.2: The 20 features most relevant in Synthetic data by RFE.

Figure 4.1 represents the accuracy of the SVM-RFE versus *t*-test, as noticed before, the accuracy for *t*-test (green) is higher, 98%, and a little bit lower for RFE (red), 96%, when the accuracy becomes constant. The accuracy distribution keeps growing while the feature size is less than 20, from 20 until the end of the features stays stable since the rest are not relevant, but noisy.

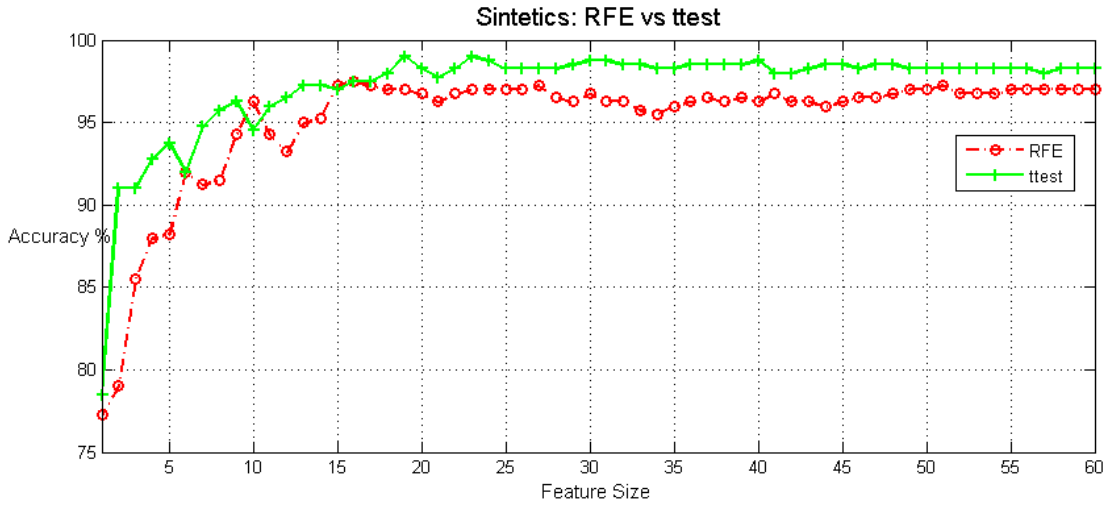


Figure 4.1: Accuracy RFE vs *t*-test for Synthetic data.

4.2 ADNI experiments

From the ADNI database we have been capable of obtaining a matrix for our algorithms of 425 samples, 195 Alzheimer patients and 230 Control subjects. The number of features representing the voxels the MRI brain scan is 521,389. All the process followed to obtain them is explained in chapter 3.

4.2.1 C value justification

First task before analyzing the experiments implementing the feature selection algorithms, SVM-RFE and t-test, with ADNI data set, is to justify the value of parameter C for SVM classifier. As mentioned previously, we have used the paper "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images" [19] as a guide and in this article small values of the penalty constant C were used. We wonder why, so we decided to make and experiment giving different values to C.

In each algorithm, when training the model with the linear classifier, a kernel type has to be set. This kernel type is the function that is going to be applied in the inner product $\langle \mathbf{x_i} \cdot \mathbf{x_j} \rangle$ of equation 2.9, also known as linear Kernel. For this experiment we can make use of a *precomputed kernel matrix* which provides an enormous computational shortcut.

The accuracy obtained for C values: 10^{-4}, 10^{-2}, 1, 10^2, 10^4. These values are assigned to a three different sizes of features, $1,000$, $10,000$ and $50,000$, as shows figure 4.2.

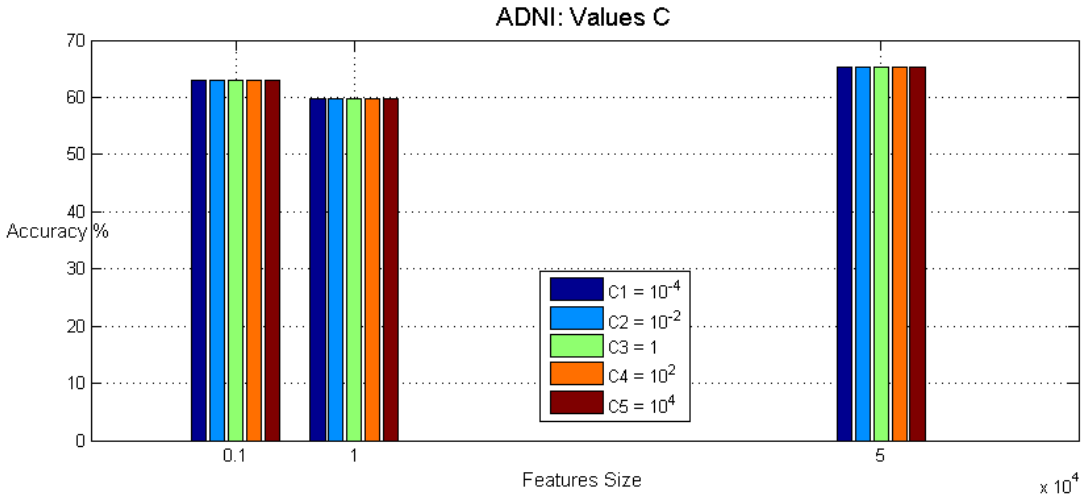


Figure 4.2: C values: 10^{-4}, 10^{-2}, 1, 10^2, 10^4 for feature size of $1,000$, $10,000$ and $50,000$ characteristics.

We are able to see that accuracy does not vary depending on the value of C. With $1,000$ features, the accuracy for each of the 5 values of C is a 63% classification error. With $10,000$, a classification error of 60% is obtained, and finally, with $50,000$ features, the accuracy has a value of 65%.

The justification of this result is that due to the high dimension difference between features ($1,000$, $10,000$ and $50,000$) and samples (425), the problem is linearly separable being the value of C irrelevant. The classifier performs so good, without misclassified samples, that a penalty constant as the C is not needed. For the executions with Synthetic and ADNI data sets the value of C has been fixed to 100.

4.2.2 Performance analysis

This subsection analyzes the accuracy of the SVM-RFE method versus the t-test one. The classification errors have been calculated for a feature size of $5,000$ to $95,000$ features, by steps of $5,000$ features. The reason of making these steps, or jumps, is justified by the big

amount of time spent by each execution, nearly 3 days for each of the feature selection methods.

When the values of the position with the most relevant features are assigned to the rank matrix, since as we have said we do not classify with the whole feature size but by steps of 5,000, we also assign values to the rank matrix in steps of 5,000. This can make a small variance in the obtained accuracy, being not fully precise, but it is needed in order to do not have an enormous computational cost.

- **SVM-RFE**

  For the SVM-RFE, as explained in subsection 2.3.3, *"Removing of the feature corresponding with the lowest value of* $|\mathbf{w_i}|$*, since is the one which the variation in the classification margin is the smallest."*, we begin training the model with 95,000 features and decreasing by steps until 5,000, measuring the accuracy in 19 points. There is a first use of the SVM-RFE algorithm, before training the classifier, filtering the best 100,000 features out of the total 521,389.

  Figure 4.3 shows that the SVM-RFE accuracy (red) stays constant in a value between 85-86% since the begining of the training model with 95,000 features until 40,000 features, in that moment the peak of the performance is reached with an exact 86.1% of accuracy. Since that moment it decreases until the 84.2%. The stability of the values in the graph, let us think that the elimination of the irrelevant/redundancy features does not improve the classification error, even it gets a little worst when features are being eliminating recursively.

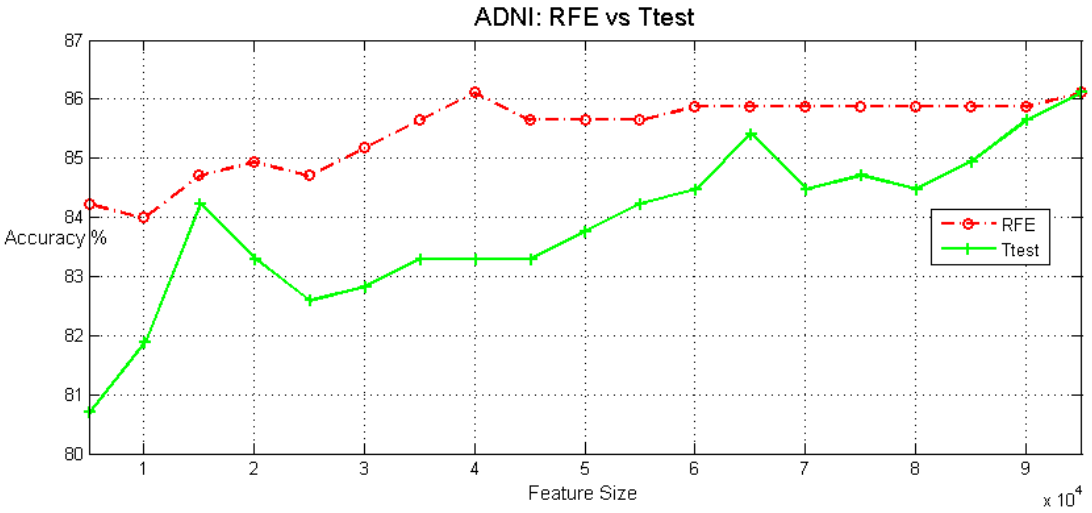


Figure 4.3: Accuracy RFE vs t-test for ADNI data.

- ***t*-test**

  In the t-test, the classifier follows a feature size steps with the same performance as the SVM-RFE, but increasing from 5,000 to 95,000, filtering, again, the best 100,000 features for the training of the model.

  Looking to figure 4.3, the t-test (green) accuracy begins with a value lower than 81%, regularly increasing, except for two peaks with 15,000 features and 65,000, until it reaches the 87% accuracy with the 95,000 features size, same as the SVM-RFE at that point.

Regardless of performing with feature selection methods, larger number of features yielded

better performance. For classifying AD and C, the feature selection achieved an average accuracy of 85.44% for SVM-RFE and 83.87% for t-test.

4.3 Visualization relevant features in MRI

In this section we will visualized the $50,000$ most relevant features of each subject which were ranked while training the SVM classifier by the feature selection methods, SVM-RFE and t-test, as it was explained in previous section.

Since we are employing a *leave-one-out* validation and test strategy (see subsection 2.4.3) over the 425 subjects that we have under study, we have obtained 425 different subsets. Each of these subsets is an array with a dimension of $521,389$ filled by ones in the positions belonging to the $50,000$ most relevant features and zeros the rest.

We are interested in obtaining a single subset, so we have merged the 425 subsets in just one, where can happen that one feature is relevant in all the iterations, in each subject, while others just in a few or in none of them.

In terms of the MRI brain scan, a relevant voxel for classification in each of the total iterations will have a value of 425, representing this the maximum intensity in the draw color range implemented to visualize these voxels in a brain template. In our case, this color range will go from red (weakest relevance) to yellow (strongest relevance).

The individual subset obtained with the relevant features is remapped to their component voxels and exported as NIfTI files, which is the standard structural MRI visualization format.

We are going to visually show the regions which belong to the relevant voxels obtained for the multivariate SVM-RFE and for the univariate t-test and analyze their color distribution in the image.

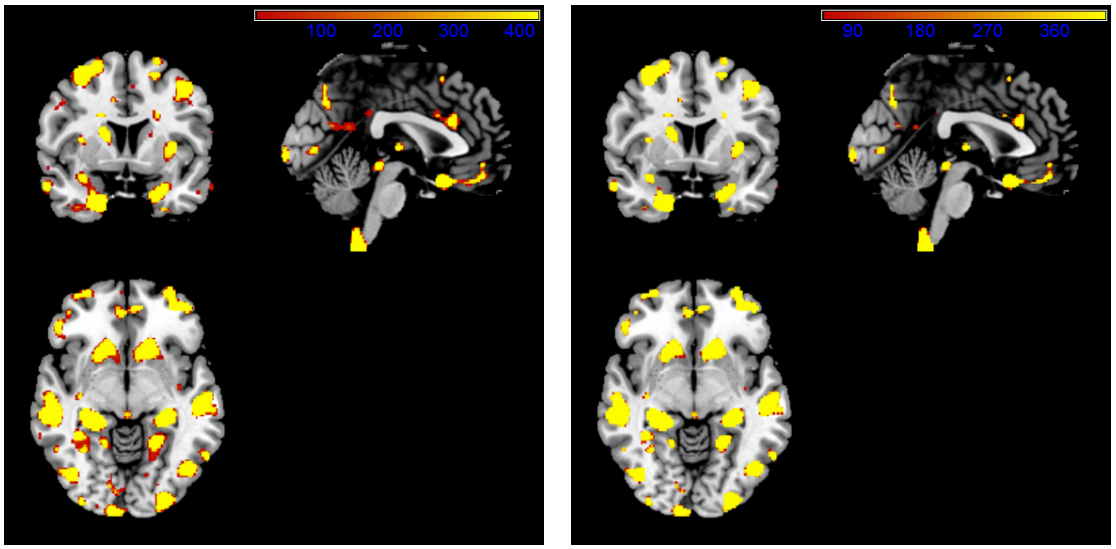
- **SVM-RFE**

  Figure 4.4a shows the selection of SVM-RFE algorithm in a 3 dimension (coronal, sagittal and axial, from right to left and from up to bottom) image centered in the anterior commissure and with their color distribution range. We see that the regions of relevance are scattered, no centered in a same area, but following a symmetric distribution between the two hemispheres of the brain.

  Looking at figure 4.5 which represents the histogram distribution of the color range, it shows that high and low values of intensity prevail, with just a few with values around 100 or 300 and none in value 200. The form was noticed when looking to figure 4.4a, since just yellow and red areas are seen, no orange ones, color that would represent those medium values of intensity of the voxels.

  As previously mentioned, the color range represents voxels with values from 1 (minimum intensity) to 425 (maximum intensity, since that voxel is relevant in all the subjects). If we set the color range in order to be 50 the minimum value, keeping out of the range the values smaller, it is noticeable in figure 4.4b, that hardly red color is presented if it is compared with figure 4.4a.

  Figure 4.7 represents the coronal, sagittal and axial views by means of longitudinal and transversal cuts in the brain. Using figure 4.6, we can make an approximation of



(a) 3-D image with minimum value of voxel 1 (b) 3-D image with minimum value of voxel 50

Figure 4.4: 3-D view of RFE-SVM agorithm representation with different values in the intensity range color (a) 1-425 (b) 50-425.

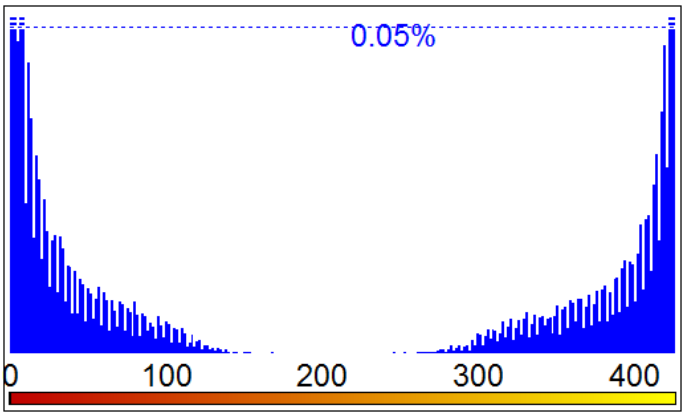


Figure 4.5: Histogram color distribution for SVM-RFE algorithm.

which regions from an axial point of view are more relevant when trying to classify Alzheimer patients from health subjects. As a first guess, going from up to bottom of figure 4.6, it seems to appear:

- Middle frontal gyrus Left and Right
- Inferior frontal left
- Cingulate region Left and Right
- Caudate nucleus Left and Right
- Superior Temporal gyrus Left and Right
- Thalamus Left and Right
- Inferior occipital gyrus Left and Right
- Occipital pole Left and Right

As mentioned earlier in this project, it is important to notice that this interpretation has to be performed by a medical expert, any other interpretation will not be valid.

The render figure 4.8 shows in a more clear way that the amount of colored voxels by the SVM-RFE is high. A total number of $77,806$ different voxels are colored.

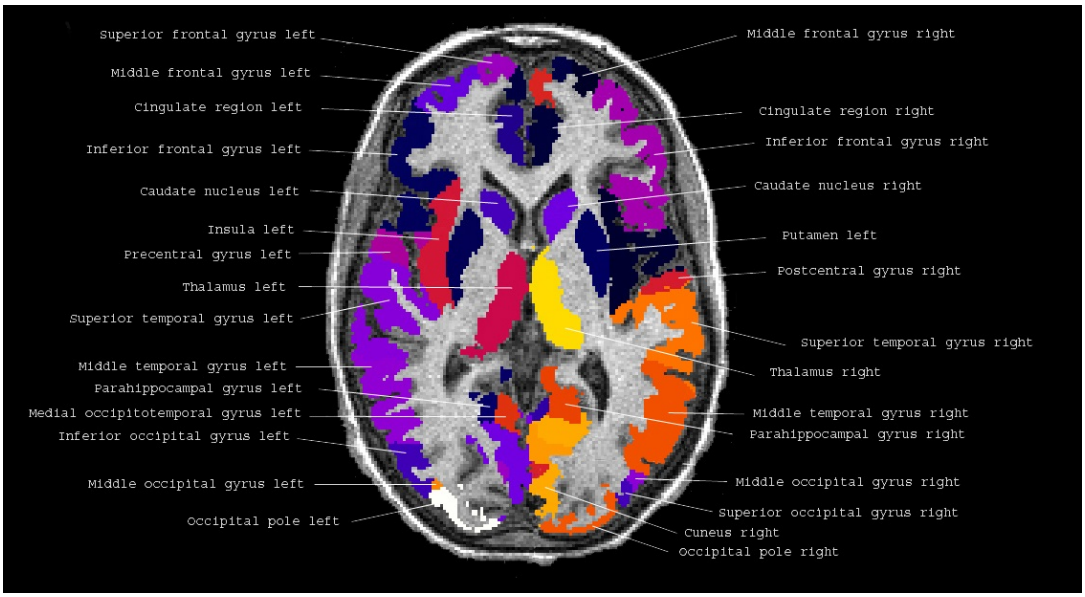
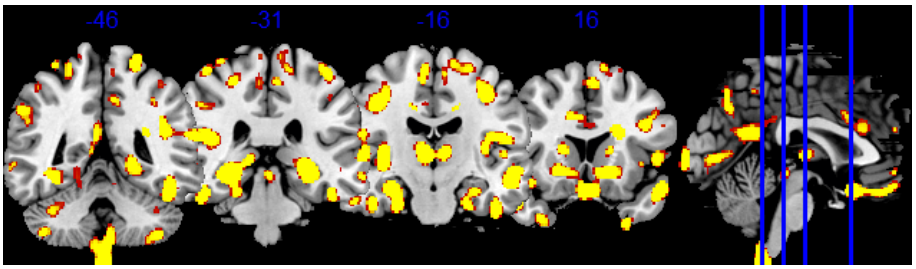
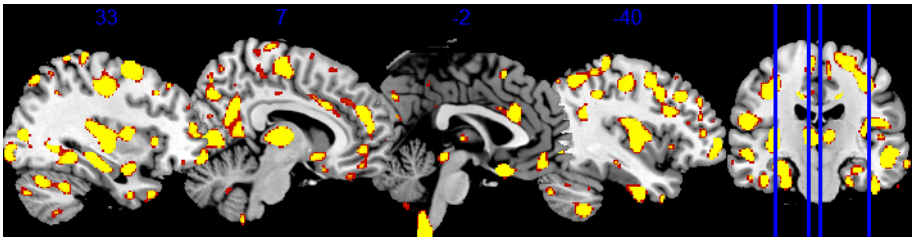


Figure 4.6: Brain regions from an axial view. Image extracted from http://www.thomaskoenig.ch/.

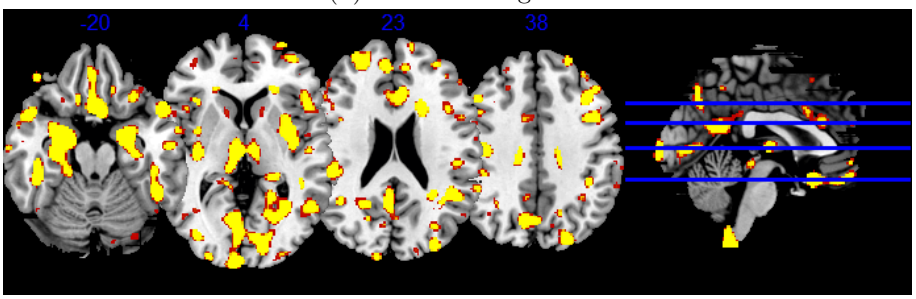
Besides, if the image penetrates to a depth of 4 mm, the brain regions of have a notorious presence of red (less relevant voxels), but if the penetration depth is 12 mm in the brain surface, the red color begins to disappear, being the yellow one more present (most relevant voxels).



(a) Multislice coronal.



(b) Multislice sagittal.



(c) Multislice axial.

Figure 4.7: Multislice representation SVM-RFEalgorithm.

- ***t*-test**

  Figure 4.9a shows *t*-test algorithm in a 3 dimension (coronal, sagittal and axial, from right to left and from up to bottom) image centered in the anterior commissure and with their color distribution range. This time the relevant voxels are located just in a few regions, not scattered in the SVM-RFE. The symmetric distribution between the two hemispheres is also presented.

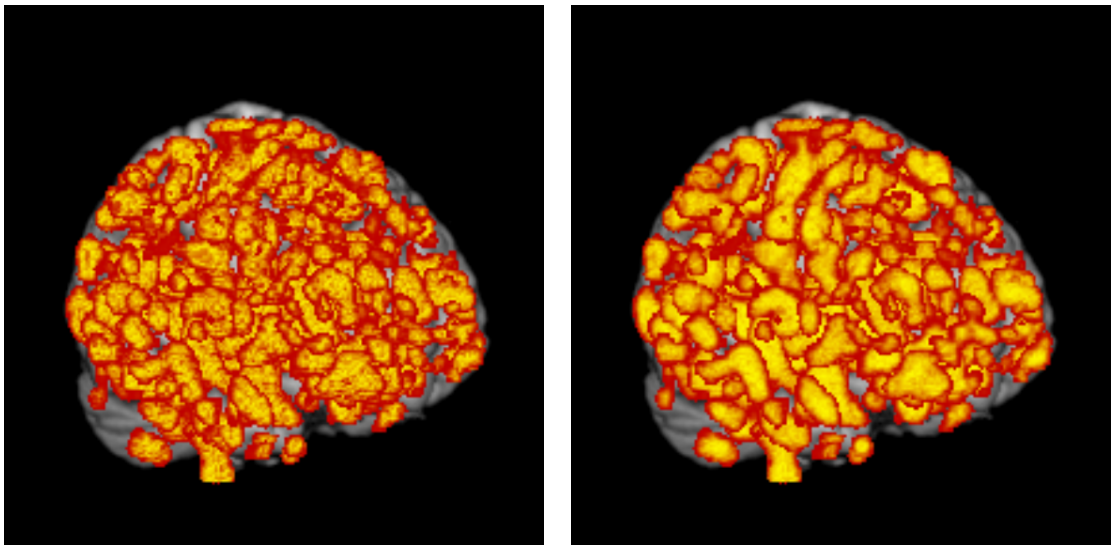
  Also represented in figure 4.10 is the histogram distribution of the color range, it shows that high and low values of intensity prevail, and just a few representing values from 50 to 350. When looking to figure 4.9a it is noticed since yellow and red areas are predominant over intermediate colors in the range.

  As previously mentioned, the color range represents voxels with values from 1 (minimum intensity) to 425 (maximum intensity, since that voxel is relevant in all the subjects). If we set the color range in order to be 50 the minimum value, keeping out of the range the values smaller, it is noticeable in figure 4.9b, that hardly red color is presented if it is compared with figure 4.9a.

  Reproducing the process follow in the SVM-RFE, using figure 4.6 and figure 4.11, we can make an approximation of which regions from an axial point of view are more relevant when trying to classify Alzheimer patients from health subjects by means of the *t*-test. As a first guess, going from up to bottom of figure 4.6, it seems to appear:

  – Superior and Middle Temporal gyrus Left and Right

  – Thalamus Left and Right



(a) 4 mm depth. (b) 12 mm depth.

Figure 4.8: Render representation for the SVM-RFE algorithm. (a) shows a depth in texture of 4 mm, while (b) the depth is of 12 mm.

– Middle occipital gyrus Right

No frontal regions presented this time, Middle frontal gyrus Left and Right, Inferior frontal left, nor Cingulate region Left and Right and Caudate nucleus Left and Right.

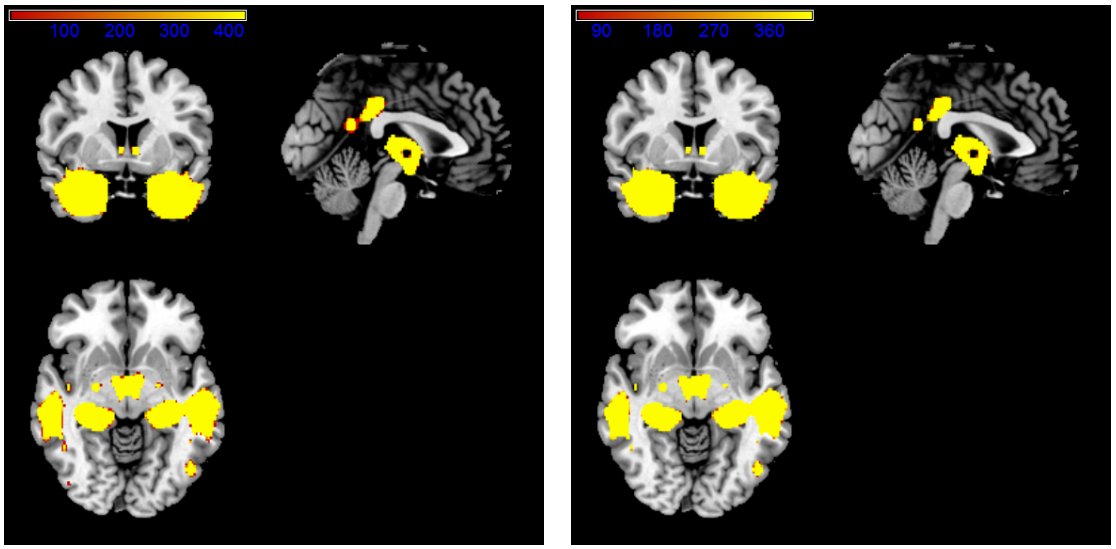
Again, as mentioned earlier in this project, it is important to notice that this interpretation has to be performed by and expert, any other interpretation will not be valid.

As a difference with respect to SVM-RFE, which relevant voxels were spread, in t-test they are concentrated in a same area, so it is difficult to guess the nucleus of importance. To deal with this, in figure 4.12 we have selected not just the $50,000$ most relevant voxels, but also the $1,000$ and $10,000$ ones, overlapping the overlays and representing them with a different color range.

The render figure 4.13 shows in a more clear way that the amount of colored voxels by the t-test. A total number of $53,670$ different voxels are colored. Compared to the $77,806$ of the SVM-RFE is quite smaller, meaning that just $3,670$ voxels are not presented in the $50,000$ most relevant voxels of each subject. Besides, if the image pass from a penetration depth of 4 mm to a 12mm one, red colored areas become weaker, being the yellow areas the predominant ones.

Comparing the visualization process of both algorithms, SVM-RFE and t-test, we have noticed that, first, the areas of relevance in the brain for each of them are different, and second, the density of voxels is much higher in the SVM-RFE than in the t-test. The reason justifying this performance was explained in chapter 2:

- In subsection 2.3.3 we explained that SVM-RFE algorithm is a multivariate testing technique, looking in a minimum subset of features, the ones that classify better, avoiding redundancy and noise. In conclusion, tries to determine which combination of variations performs the best out of all of the possible combinations.
- In the case of the t-test, as it is a univariate analysis algorithm, exploring each variable in a data set separately, being probable that keeps redundancy variables, having as a consequence that more connected regions appear in the visualization.



(a) 3-D image with minimum value of voxel 1. (b) 3-D image with minimum value of voxel 50.

Figure 4.9: 3-D view of *t*-test agorithm representation with different values in the intensity range colour (a) 1-425 (b) 50-425.

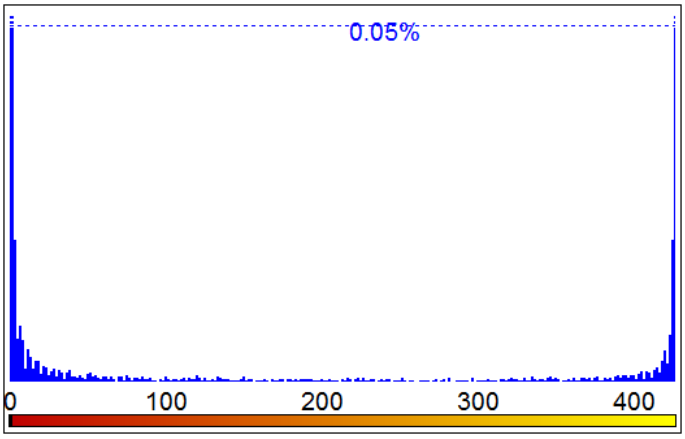
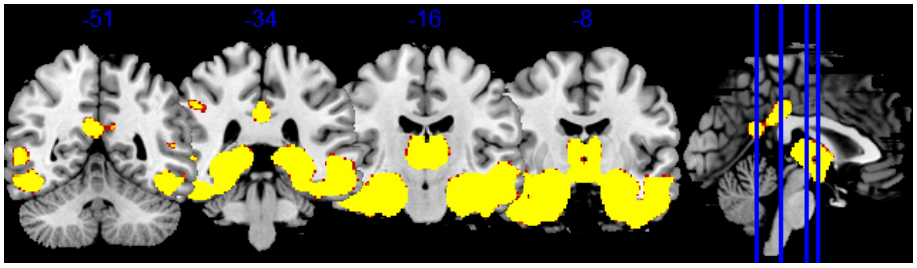
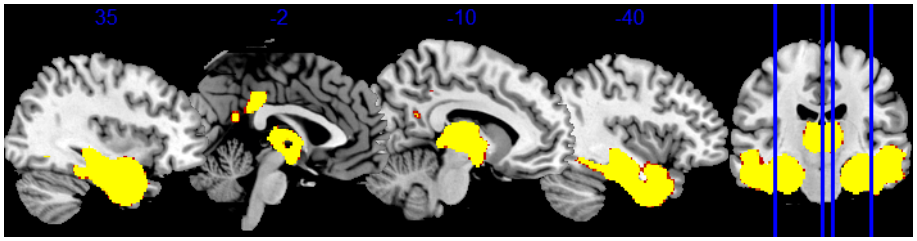


Figure 4.10: Histogram color distribution for *t*-test algorithm.

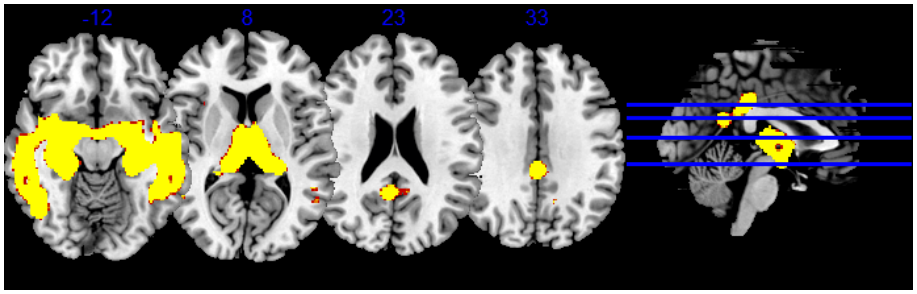
For a problem of interpretation like this one, it is important to know everything that is relevant, even if it redundant. So *t*-test would be the most indicated for the automatic selection of features for Alzheimer detection. But as repeated several times, is an expert in the medical field who has the ability for analyzing it properly.



(a) Multislice coronal.



(b) Multislice sagittal.



(c) Multislice axial.

Figure 4.11: Multislice representation t-test.

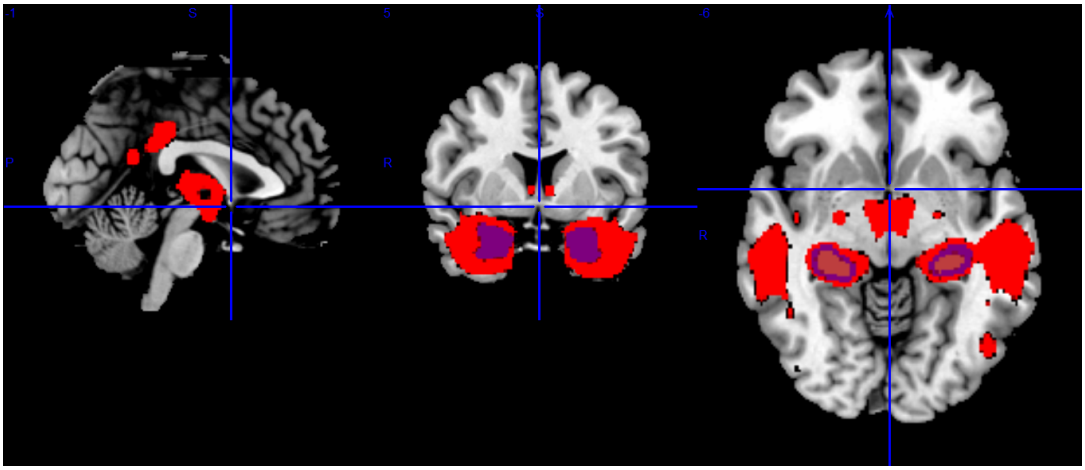
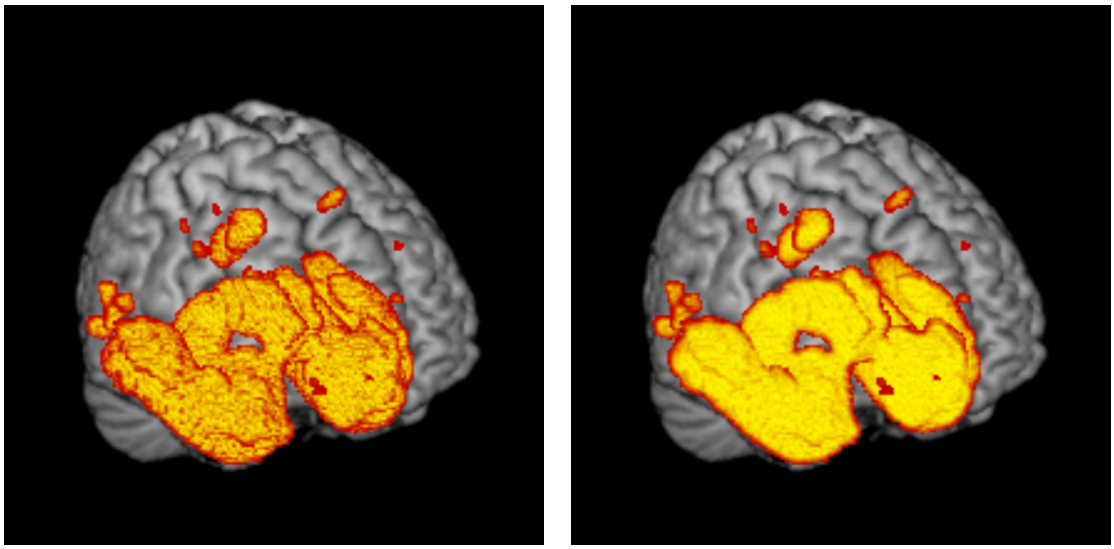


Figure 4.12: Representation of the most $1,000$ (cyan), $10,000$ (violet) and $50,000$ (red) relevant voxels of the t-test algorithm.



(a) 4 mm depth. (b) 12 mm depth.

Figure 4.13: Render representation for the t-test algorithm. (a) shows a depth in texture of 4 mm, while (b) the depth is of 12 mm.

Chapter 5

Conclusions and future lines of investigation

5.1 Project conclusions

Several conclusions can be extracted from the study, implementation, testing and visualization processes carried out along this project. Besides, the achievement grade of the objectives presented at the beginning, in section 1.2, can now be evaluated.

- From previous study of the state of art, it is a common drawback the lack of samples in feature selection methods for Alzheimer detection. Due to the high difficulty of getting MRI brain scans with similar imaging characteristics for making a precise classification. Even comparing with the paper [19] we had as a guide, we were able to get in this project 106 useful and good quality MRI brain scans more than them. By an intensive examination of the ADNI database and a precise process of preprocessing, the achievement of this objective was possible.

- Once the preprocessed of the MRI brain scans was finished, a $521,389$ voxels image was obtained, being then vectorized to a matrix with $521,389$ features as columns to perform a classification between Alzheimer and Control subjects. This dimensionality was huge, impossible to make any of our implementations with the resources we had at hand due to the extremely high computational cost that it would take. By the application of feature selection methods, RFE and t-test, a huge reduction was possible, being capable of doing a first filter for getting the best $100,000$ out of the total features. And then classifying the model by steps of $5,000$, drastically reducing the computational cost. Being the proof of a great result, but being desired a highest reduction. Remark that part of this great result is a consequence of the knowledge gain and the trails done by the Synthetic data.

- Our project has been able to classify between Alzheimer and Control subjects successfully. Using the *libsvm* MATLAB toolbox as the linear classifier to distinguish between both populations we have reached average accuracies of 85.44% for SVM-RFE and 83.87% for t-test. We can conclude that our classifier performs in a high level of accuracy.

- A deep evaluation and its corresponding clear explanation and interpretation has been done. The graphics showing the results of the accuracies and the justification

of the parameters value used during the implementation. Can be considered as a complete global analysis of the project.

In conclusion, we can affirm that most of the objectives have been reached, anyway, dimensionality still being high and the accuracy could be improved, they are positive and can consider as a great first step for future feature selection projects taking advantage of these preprocessed MRI scans and the algorithms implemented. It is needed to be remarked that all these results have to be seen as a complementary medical information for a diagnosis, being mandatory to be analyzed by an expert in the field, but indeed, our work could provide society with a very powerful and useful tool in the diagnosis and treatment of mental illness.

5.2 Future lines of investigation

This project has many possibilities for being improved, using different techniques for the implementation or even being applicable in a different context. Although we have followed previous studies, the different combinations of algorithms and use of different data sets, gives the possibility of future lines of investigation.

We have divided this possible future research into: Improvement of detection process and use of new feature selection methods.

5.2.1 Improvement of detection process

The obtained results are considered quite good, but many new functionalities can be implemented and new resources used. There are some of them that we consider that could be important:

- The ADNI initiative, during the whole process of the Alzheimer disease, stores not just MRI of healthy or ill patients, but also intermediate phases as the early mild cognitive impairment (EMCI) and late mild cognitive impairment (LMCI). The samples belonging to those phases could be also preprocessed and select their most relevant features in order to see a temporal evolution of the disease process. We could check the change of the relevance of the features while the diagnostic advances from a healthy subject to an Alzheimer patient. For its implementation, it would be needed to use a Multiclass SVM linear classifier, since it would not be anymore a binary problem, but a multiple class problem.
- More characteristics could be analyzed when selecting the most relevant features in the MRI. We have available the gender and age of the samples, and the analysis of the possible correlation between groups of different ages, males and females could give the change of measuring how the relevant features variate.

5.2.2 Use of new feature selection methods and linear classifiers

We employed the SVM linear classifier and the SVM-RFE and t-test feature selection methods, but as mentioned in chapter 2, there are more options. The use of this classifier and methods were previously justified due to the positive performance proved in

previous studies, but to compare between different options would provide a big amount of information.

- Linear classifiers Perceptron and Fisher's implementation with the ADNI database would give the chance to compare their performance to the one implemented, to justify the use of the SVM with respect to the others.
- Several studies use the regions of interest (ROIs) on brains to calculate the enclosed volume. This process preselect some regions in the brain with high probability of storing relevant voxels for the Alzheimer detection, as could be the hippocampus and parahippocampal, focusing in these specified regions instead on all the brain. However this process tends to be complex and time consuming, being the reason of why we did not implement it, but it would be a valuable experiment.

APPENDICES

APPENDIX A

Project planning

This **Appendix A** explains the structure of the project's planning. This project started the 15^{th} of June of 2015 and finished the 15^{th} of February of 2016.

The structure consists on several stages, each of them identified by the tasks carried out but with the common characteristic of being supported by a continuous contribution of the tutor's knowledge and help, and the always presented necessity of the student to search in all the available documentation related with the project's issue.

The initial stage consisting on an introduction of the project and a preliminary scope by the examination of the data available in the ADNI initiative.

Once the objectives were established, the long process of downloading data samples and learning the required knowledge took place, being followed by the filtering of the MRI images by means of the development of Java scripts to get the ones that would be preprocessed.

Preprocessing was an intense phase due to the manual reorientation and the long process of segmentation and normalization of MRI images.

Then, we focused on classification and its widely range of possibilities. A practice of training by means of Synthetic data made up by us was done to get used to the classification algorithm.

Afterwards, we performed the implementation of feature selection algorithms.

Finally, once the results were considered optimum, it was time for writing the bachelor memory.

A concrete separation of the tasks, grouped in 5 phases, and their duration is shown in table A.1. Also, a representation of a Gantt graph is shown in figure A.1.

Task	*Start*	*Duration (days)*	*End*
Initial phase			
Project introduction	*15/06/15*	*1*	*15/06/15*
ADNI access request	*16/06/15*	*6*	*21/06/15*
Available data examination	*22/06/15*	*3*	*24/06/15*
Objective's statement	*25/06/15*	*1*	*25/06/15*
Data and knowledge acquisition & Preprocessing			
Data download	*26/06/15*	*32*	*27/07/15*
Conceptual transfer	*26/06/15*	*32*	*27/07/15*
Tool familiarization	*28/07/15*	*9*	*05/08/15*
Vacation	*06/08/15*	*9*	*14/08/15*
Implementation filter data scripts	*15/08/15*	*15*	*29/08/15*
Selection final MRI samples	*30/08/15*	*6*	*04/09/15*
Study of previous work	*05/09/15*	*10*	*14/09/15*
Re-orientation MRI images	*15/09/15*	*20*	*04/10/15*
Segmentation & Normalization process	*05/10/15*	*16*	*20/10/15*
SVM Classifier & Feature Selection algorithms			
Study SVM Classifier	*21/10/15*	*11*	*31/10/15*
Training synthetic data	*01/11/15*	*7*	*07/11/15*
T-test implementation	*08/11/15*	*23*	*30/11/15*
RFE implementation	*01/12/15*	*20*	*20/12/15*
Results validation			
Results visualization	*21/12/15*	*3*	*23/12/15*
Performance analysis	*24/12/15*	*17*	*09/01/16*
Final phase			
Memory	*10/01/16*	*36*	*14/02/16*

Table A.1: Project task list and the corresponding dates of duration. The project started the 15^{th} of June, 2015.

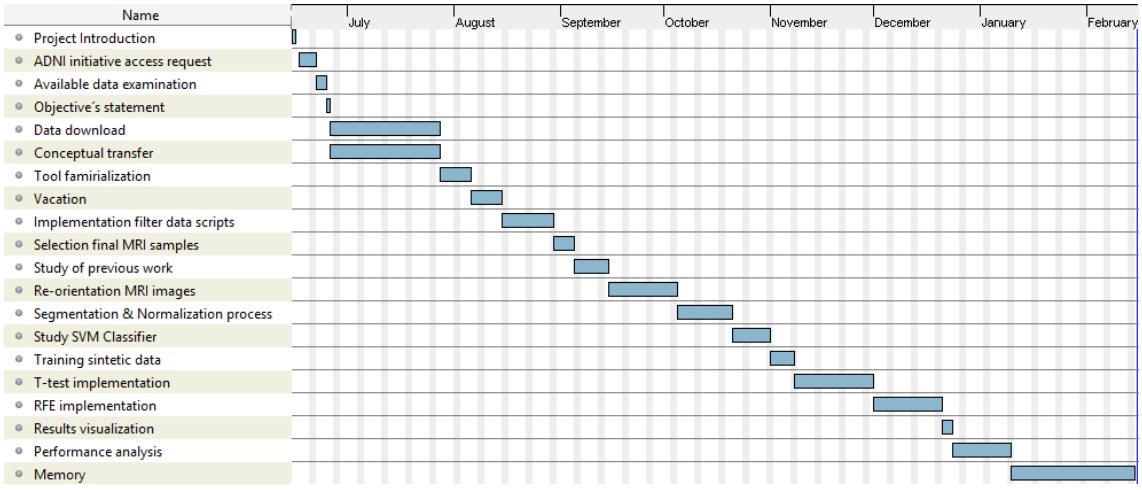


Figure A.1: Project Gantt graph.

APPENDIX B

Project budget

This **Appendix B** presents the global costs for implementing this Bachelor Thesis. The global costs are a set of expenditures made up by the personnel, material resources and indirect costs.

- **Personnel costs**

  Two persons have been working in this project. A student, considered as a junior engineer, and a tutor, working as a senior engineer. Besides, this project is based on 5 stages, as shown in B.1, and the sum of the time spent by student for each phase sums a total number of hours of 900 hours. The cost of these working hours by the student and the ones of the tutor, which were spread along the whole project, are calculated in table B.2, having a total cost of 15,600 euros.

Project Stages		
Stage 1	*Initial phase*	70 *hours*
Stage 2	*Data and knowledge acquisition & preprocessing*	330 *hours*
Stage 3	*SVM Classifier & Feature Selection algorithms*	250 *hours*
Stage 4	*Results validation*	100 *hours*
Stage 5	*Memory*	150 *hours*
Total		900 *horas*

Table B.1: *Project Stages.*

Personnel Costs				
Name	***Categories***	***Working hours***	***Cost euro/hour***	***Cost (€)***
Jesús Fernández Bes	*Senior Engineer*	80	60	4,800
Nuño de la Concha Vega	*Junior Engineer*	900	20	10,800
Total				15,600

Table B.2: *Personnel costs.*

- **Material resources costs**

  For the implementation of this project, several material resources have been needed. Resources like informatics equipment, licenses, MRI brain scans, ... among others, are grouped in table B.3, summing a total cost of 2,310.7 euros.

Concept	Quantity	Price (euros/unit)	Amort.% (yearly)	Total
MRI brain scan	425	1,000	0[1]	
External disk 2TB	1	75	100	75
Individual Matlab license	1	2,000	100	2,000
PC1 (personal)	1	500	40	200
PC2 (university machine)	1	7,000	0.51[2]	35.7
Total				2,310.7

Table B.3: *Material resources costs*

[1] ADNI initiative has had 10,435,975 downloads and 80,661 users have taking advantage of this non-profit association, for this reason the ADNI MRI brain scans are free cost.

[2] Assuming 4 years of life and 22 hours of daily use. Then calculating the executions duration implemented have taken 150 hours, an amortization of 0.51% is obtained.

- **Total project budget**

  By adding the whole set of previous calculated costs, the overall budget is shown in table B.4.

Total Cost	
Concept	***Cost***
Personnel costs	15,600 €
Material resources costs	2,310.7 €
VAT (21%)	3,761.24 €
TOTAL	21,671.95 €

Table B.4: *Total budget.*

The total budget of this project is 21,671.95 euros.

Madrid, 24^{th} of February 2016

The project Engineer

Nuño de la Concha Vega

APPENDIX C

Summary

This **Appendix C** is a summary of the development of the project "Automatic Selection of features in MRI for Alzheimer detection".

The motivation for carrying out this project has to be with the huge social impact of Alzheimer in nowadays society. Alzheimer has become one of the most disturbing and dangerous illness, being, in terms of damage, in the top of the most harmful diseases. As a neurodegenerative disease, appears by the deterioration the cognitive and behavioral abilities of the superior brain functions. Memory loss, lack of orientation, problems for communication, among others, are the serious consequences that patients suffer. Patients are not the only ones facing the consequences of the disease, patient's families have to handle not just with this social impact, but also with the economical one. Even though Alzheimer has become a big cost for governments budget in terms of sanity, a big percentage of the expenses are taken by patient's family.

This growth of awareness in the society with respect to this mental disorder have pushed the development and implementation of new techniques to come out with an earlier diagnosis that can reduce the impact of the disease and maybe, in a future, its cure.

When diagnosing, apart from the behavioral assessments, new neuroimaging analysis techniques, as a complement for the behavioral assessments, have been employed during the diagnosis process last years. Techniques as Multivariate Pattern Analysis (MVPA) have been used to study the patterns of neurodegenerative diseases and mental disorders using Magnetic Resonance Imaging (MRI). The analysis of structural MRI brain scans for the different neurodegenerative diseases have been employed in neurofunctional techniques, as Voxel-Based Morphometry (VBM) giving the capability of researching for the possible differences in the brain anatomy.

Not only the techniques in themselves, but also the algorithms to implement them have been improved and its employment has increased. Machine learning (ML) as a subfield of the computer science, has been used in evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions.

We have implemented all the methods and techniques in this project with the principal objective of the automatic selection of features in MRI to detect Alzheimer. We have

divided this main objective into a group of subobjectives:

1. Selection of samples set of MRI brain scans with common imaging characteristics through the access and examination of the Alzheimer´s Disease Neuroimaging Initiative database. Following pre-established patterns of searching and downloading, get as many Alzheimer and control subjects as available in the database. Preprocessing of these MRI brain scans to obtain the features from the voxels of the MRI.
2. Application of feature selection techniques to reduce the high dimensionality of features on each MRI. Implementation of the RFE and t-test methods to eliminate redundant and irrelevant features, avoiding noise and difficult interpretation, and keeping the most relevant ones. Use of synthetics data samples generated by us for a more clear understanding and easier development.
3. Use of a linear classification to distinguish between healthy and ill subjects. A linear classifier SVM will be implemented and MATLAB toolboxes used for the visualization of the MRI and the development of the algorithms. Try to get the highest accuracy possible in each of the methods.
4. Evaluation of the performances of each of the implemented techniques. An analysis of the different strategies followed by each method and their classification error in the experiments. We are interesting in getting a final visualization of the most relevant features in its context, a brain template.

After setting the objectives of the project, we have introduced the machine learning problem as a classification problem. In such a problem, our objective is to predict the class of some data samples out of a discrete number of known classes.

In a first step in the learning process a large set of N inputs $\{x_1, ..., x_N\}$, called a training set, is used to set the parameters of an adaptive model by a Machine learning algorithm. This training set has been previously classified and labelled in their respective categories. The machine learning algorithm is run, which can be understood as having a function *y=f(x)*, where input *x*, training set, is used to generate an output textity, labels. This process corresponds to the training or learning phase.

To the model training follows the testing phase, where new data inputs, completely different to the ones corresponding to the training set, are used to measure the accuracy of the model generated. The ability of categorizing this new data set is known as generalization. This ability can be consider the main goal of this process since is needed to build a successful model should be able to classify different inputs, not just fit to the training ones, avoiding the risk of overfitting.

This is exactly what we wanted to achieve, to identify the class of a subject, Alzheimer or Control, by means of their features. To the end of this application, the algorithm that has been used, produces a dependency between the inputs and the desired outputs of the system. This dependency is obtained through the learning acquired thanks to previous labelled samples training. This description corresponds to the algorithm of the supervised learning, one of the types of machine learning algorithms which follow a taxonomy in function of the generated outputs by the system.

The classification of the subjects, based on the class they belong to, has been obtained by a linear classifier through a lineal combination of their characteristics. The characteristics, also known as features or even Neuromarkers for this case, will be real numbers which corresponds to the voxels of the MRI images.

Our study contained a data set with an extremely big number of dimensions, more than 500, 000 elements, representing the voxels of the MRI images, and a quite small number of subjects, more than 400, meaning that our problem was separable; the ratio of possible decision thresholds is infinite and all of them corrects.

Although there are multiple algorithms that solve the required problem, like the Perceptron or Fisher´s classifiers, the Support Vectors Machines classifier has been the one used in this work, since our main objective was not to classify but to make a selection of the most relevant features of the MRI images, the ones which provide the highest level of information to classify subjects. The SVM assigns to each feature a weight according to its relevance.

Support Vectors Machines stand out not just but their capacity of generalizing the model, but also because of their capability for minimizing overfitting in the training set, capability that is not presented in other learning methods like the neural networks, all of these justified reasons for their use.

This SVM model builds up a hyperplane or a set of hyperplanes in a space of high dimensionality (even infinite). Infinite discriminatory hyperplanes exist to divide the samples, because of this, is desired the one which classify in the most optimum form. This optimum one has the highest margin of separation between samples of different classes, reason of why the SVM are also known as "classifiers of maximum margin". Following this, samples labelled in one class will be on one side of the hyperplane, and samples belonging to the other class, in the opposite side.

Classification by SVM can be carried out by a linear or non-linear separation in the surface of the samples, always with the main object of minimizing the error classification, reaching for the highest accuracy.

We have performed for the linearly separable case, valid for a binary classification problem, in our case Alzheimer patients (AD) versus Control subjects (C), but all the obtained conclusions in this binary case can be extrapolated to a Multiclass SVM, generating as many classifiers as classes available.

In order to solve the classification problems, using machine learning methods, the issue of the large number of input variables presented in the data had to be faced. Working with the MRI images, the number of variables, in this case the number of voxels, was extremely high, more than 500, 000, quite larger than the number of subjects, 425.

This extremely high number of variables would have caused some complications: complex models, which would have been hard to visualize and to understand by researchers, high measurement and storage requirements would have been needed, increment of training and utilization times and presence of redundant and/or irrelevant features, adding noise rather than providing useful information for the classification. All these drawbacks have been avoided by means of feature selection.

A feature selection algorithm can be seen as the combination of search techniques for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets and find which of these subsets of features is the most relevant and informative. Out of these feature selection methods, we have implemented the ***t*-test** and the **SVM-RFE**.

- ***t*-test**

  "The independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared".

  This univariate analysis explored each variable in a data set, separately. It looked at the range of values, as well as the central tendency of the values, describing the pattern of response of each variable on its own, individually.

  It is commonly applied when the statistical units underlying of the two samples being compared are non-overlapping. From the point of view of the scenario of this project, it was wanted to identify if a patient suffered of Alzheimer or was a Control subject, for that, more than 400 subjects were enrolled, approximately 200 subjects were assigned to the patients group and 200 subjects to the control group. In this case, there were two sets of independent samples and would be used the unpaired form of the two-sample t-test.

  Applying the hypothesis that the data from the two populations come from independent random samples from normal distributions with equal means and equal, but unknown, variances. This is known as the null hypothesis and is said to be accepted. The alternative hypothesis is that each population has a different mean. The function returns the result of the test in H. If the value obtained of H is 0, the null hypothesis is true, equal means. Otherwise, if H is equal to 1, corresponds to a small p-value, which casts doubt on the validity of the null hypothesis, being rejected at the 5% significance level.

  As it was desired to find between all the features, the ones which distinguished between both populations, the ranked list of features was based on the p-value of the t-test, being the features with small p-values at the top for the characterization between Alzheimer patients and Control subjects

- **SVM-RFE**

  The Recursive Feature Elimination was implemented as an iterative feature selection algorithm developed specifically for Support Vectors Machines. The RFE algorithm has selected features depending on the classification margin supplied by the classifier of the SVM, there was not a separation between the learning process and the feature selection one.

  The SVM-RFE represents a multivariate testing technique for testing a hypothesis in which multiple variables are modified. The goal of multivariate testing is to determine which combination of variations performs the best out of all of the possible combinations.

For this project, it was implemented a first iteration used to select the best $100,000$ variables out of the more than $500,000$, eliminating the rest, which corresponds with the lowest weight value. Then, during the training, every iteration would rank and eliminate variables in steps of $5,000$ variables. Despite the fact that performing in this way could reduce the accuracy of the classification, these methods have provided a classification of the subsets of samples that have finished faster than eliminating individually.

As important as the training methods, was the performance evaluation. One usual performance measurement for classification problems is the 0/1 loss function, i.e., percentage of misclassified samples. However, it is very important to remark that our objective was not to minimize this cost over the training data but to evaluate how good our approach classified test samples, i.e. how well our method generalizes to new data. Consequently

we had to use some technique that allowed us to approximate that performance from the labeled data we had.

There are two accuracy estimation methods which are the most common ones, **Conventional Validation** and **Cross-Validation**. In this thesis we have used the **Cross-Validation**.

This technique evaluates the statistical analysis results and guarantees the independency between training and test data. The two subsets of data are used to perform the cross-validation analysis, which is the arithmetic mean of the obtained evaluation measures over different partitions, but with the difference that many rounds, changing the data belonging to the partitions, are executed, getting a reduction of the variance and being the validation results the average over the iterations. This behavior enables a more accurate estimation of the model prediction performance.

In our project the *leave-one-out* cross-validation (**LOOCV**) was the performance evaluation employed. The reason is that it maximizes the amount of available data for training getting a better accuracy of the SVM classifier, despite the computational cost.

In practice, the way in which was performed the **LOOCV** was taking the whole amount of subjects we had available, 425, and take one different sample in each iteration as validation sample and the rest for training.

Then, the SVM linear classifier proceeded to optimize the model parameters to make the model fit the training data the best as possible, then the generalization of the model would be done by the independent samples of the validation subset, just one sample in our case.

before the training process, it was carried out the data acquisition and image preprocessing.

The input data space employed in this project has been acquired from the Alzheimer's Disease Neuroimaging Initiative database.

Out of all the available clinical data and subjects belonging to all the phases of the Alzheimer evolution process, we have chosen the MRI images of the Alzheimer and Control subjects. We decided to select these two classes for the study instead of the also available EMCI (Early Mild Cognitive Impairment) and LMCI (Late Mild Cognitive Impairment) to limit the scope of this project and for considering them more clearly differentiable in the classification.

In this part of the project we have focused on selecting the highest number of samples with the same imaging characteristics, acquiring a similar number of AD and Control, and following the preprocessed methods needed to optimally prepare the data samples for the training model in order to obtain the highest accuracy.

Since our objective was to find the biomarkers of the brain regions that are most relevant to Alzheimer, we employed the structural Magnetic resonance imaging (sMRI), since it provides anatomical reference for visualization of activation patterns and regions of interest in the brain to extract functional signal information from Alzheimer and Control subjects.

The MRI, is a medical imaging technique used in radiology to image the anatomy and the physiological processes of the body in both health and disease. MRI scanners use strong magnetic fields, radio waves, and field gradients to form images of the body. The structural MRI provides information that is used to describe the shape, size and integrity

of different tissues in the brain.

The main reason of using the structural MRI images was that are highly sensitive for detecting relevant neurodegenerative changes in Alzheimer disease and are used as an outcome measure for clinical trials.

The neuroimages we have used from the ADNI database are basically the T1-weighted anatomical MRI brain scans in 3-D with a field strength of 1.5 Tesla. The reason of using the T1-weighted MRI was that "provides a good contrast between gray matter (dark gray) and white matter (lighter gray)" because are the regions in th brain where the loss and deterioration of the superior brain functions consequence of Alzheimer's disease are easier to appreciate.

All of these MRI images has been downloaded in a Neuroimaging Informatics Technology Initiative (NIfTI) format from the ADNI database.

Once the download have finished, we have obtained 195 AD samples and 230 Control samples, being a total set of 425 samples.

For this project we took as a reference the paper "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images". The reason is because the statistical analysis they have carried out are similar to what we have planned to do and they have used the Alzheimer's samples from the database as well.

MATLAB (**MAT**rix **LAB**oratory) was the mathematic software tool that we employed to develop each of the experiment scripts and visualize images and figures. It was chosen due to its capability for matrix manipulation with a high precision of calculation, reliable representation of data, functions and algorithms implementation and its big number of available toolboxes.

These images were preprocessed to transform them into a new space of variables in order to solve in an easier way the pattern recognition problem. The preprocessing process was implemented with the Statistical Parametric Mapping (SPM12). The SPM12 is an academic software toolkit for used for the analysis of functional imaging data.

The preprocessed process consisted on: **Re-orientation**, **Segmentation** and **Normalization**.

Once the preprocessed and the implementation of the algorithms phases were over, we analyzed the performance.

The analysis of the performance of each of the feature selection strategies tried to determine the most effective combination. It also showed the evolution of the classification test error with respect to the number of selected features.

All features selection algorithms implemented were previously designed and executed with a sintetic data set made up by ourselves. It was a set that simulated the characteristics of the ADNI one but easier to handle. The reason of using it was due to the high dimensionality, $521,389$ features after the vectorization, of the ADNI MRI images obtained from the preprocessing process.

In the analysis of algorithms for t-test and SVM-RFE, we payed special attention to the

results obtain from a *rank order* vector which stored the ranking of all the features of each subject. This allowed to see in which positions are located the best features for the classification of the samples. Besides, the accuracy % which mesuares the success of the classification.

After training and analyzing the results of the synthetic data, it was time for the ADNI dataset.

First task before analyzing the experiments implementing the feature selection algorithms, SVM-RFE and *t*-test, with ADNI data set, was to justify the value of parameter C for SVM classifier. We wonder what value we should use, so we decided to make and experiment giving different values to C.

We were able to see that accuracy did not vary depending on the value of C, so we fixed the value to 100. As the problem was linearly separable, the value of C was irrelevant. The classifier performed so good, without misclassified samples, that a penalty constant as the C was not needed.

Focus on ADNI dataset, the classification errors were calculated for a feature size of $5,000$ to $95,000$ features, by steps of $5,000$ features. The reason of making these steps, or jumps, was justified by the big amount of time spent by each execution, nearly 3 days for each of the feature selection methods.

When the values of the position with the most relevant features were assigned to the rank matrix, we also assigned values to the rank matrix in steps of $5,000$. This could make a small variance in the obtained accuracy, being not fully precise, but it was needed in order to do not have an enormous computational cost.

At the end, regardless of feature selection methods, larger sample sizes yielded better performance. For classifying AD and C, the feature selection achieved an average accuracy of 85.44% for SVM-RFE and 83.87% for *t*-test.

After obtaining the training results, we visualized the $50,000$ most relevant features of each subject which were ranked while training the SVM classifier by the feature selection methods, SVM-RFE and *t*-test.

Since we were employing a *leave-one-out* validation and test strategy over the 425 subjects that we had under study, we obtained 425 different subsets. Each of these subsets is an array with a dimension of $521,389$ filled by ones in the positions belonging to the $50,000$ most relevant features and zeros the rest.

We were interested in obtaining a single subset, so we merged the 425 subsets in just one, where could happen that one feature was relevant in all the iterations, in each subject, while others just in a few or in none of them.

Comparing the visualization process of both algorithms, SVM-RFE and *t*-test, we noticed that, first, the areas of relevance in the brain for each of them were different, and second, the density of voxels was much higher in the SVM-RFE than in the *t*-test. The reason was explained in the theory part.

- SVM-RFE algorithm is a multivariate testing technique, looking in a minimum subset of features, the ones that classify better, avoiding redundancy and noise. So, it tries to determine which combination of variations performs the best out of all of

the possible combinations.

- In the case of the t-test, as it is a univariate analysis algorithm, exploring each variable in a data set separately, being probable that keeps redundancy variables, having as a consequence that more connected regions appear in the visualization.

For a problem of interpretation like this one, it was important to know everything is relevant, even if it redundant. So t-test was the most indicated for the the automatic selection of features for Alzheimer detection. But as repeated several times, is an expert in the medical field who has the ability for analyzing it properly.

From this analysis performance, several conclusions were extracted from the study, implementation, testing and visualization processes carried out along this project. Besides, the achievement grade of the objectives presented at the beginning, could be evaluated.

- From previous study of the state of art, it was a common drawback the lack of samples in feature selection methods for Alzheimer detection. Due to the high difficulty of getting MRI brain scans with similar imaging characteristics for making a precised classification. We were able to get in this project 425 useful and good quality MRI brain scans, despite of our limitations of resources and knowledge as a university thesis bachelor project. By an intensive examination of the ADNI database and a precise process of preprocessing, the achievement of this objective was possible.

- Once the preprocessed of the MRI brain scans was finished, a $521,389$ voxels image was obtained, being then vectorized to a matrix with $521,389$ features as columns to perform a classification between Alzheimer and Control subjects. This dimensionality was huge, impossible to make any of our implementations with the resources we had due to the extremely high computational cost that it would take. By the application of feature selection methods, RFE and t-test, a huge reduction was possible, being capable of doing a first filter for getting the best $100,000$ out of the total features. And then classifying the model by steps of $5,000$, highly reducing the computational cost. Being the proof of a great result, but being desired a highest reduction. Remark that part of this great result was a consequence of the knowledge gain and the trails done by the synthetic data.

- Our project was able to classify between Alzheimer and Control subjects successfully. Using the *libsvm* MATLAB toolbox as the linear classifier to distinguish between both populations we had reached average accuracies of 85.44% for SVM-RFE and 83.87% for t-test. We could conclude that our classifier performed in a high level of accuracy.

- A deep evaluation and its corresponding clear explanation and interpretation was done. The graphics showing the results of the accuracies and the justification of the parameters value used during the implementation. Can be considered as a complete global analysis of the project.

Then the structure of the project's planning was explained. This project started the 15^{th} of June of 2015 and finished the 15^{th} of February of 2016. And a deep explanation of the implemented tasks and their duration was explained.

Also the budget of the project, with its personnel, material resources and overall cost was explained in detail, with a total cost of $21,671.95$ euros.

This bachelor thesis was carried out by Jesús Fernández Bes, as Senior Engineer, and Nuño de la Concha Vega, as Junior Engineer.

Bibliography

[1] California state university long beach. `http://www.csulb.edu/`.

[2] Mathworks official web page. `http://mathworks.com`.

[3] Nifti in mathworks. `http://www.mathworks.com/matlabcentral/fileexchange/8797-tools-for-nifti-and-analyze-image`.

[4] University of california, san diego. `http://fmri.ucsd.edu/`.

[5] Alzheimer's disease neuroimaging initiative (adni). `http://adni.loni.usc.edu/`, 2005.

[6] Introduction of adni study basics. `http://adni-info.org/`, 2005.

[7] Mr contrast agents. `http://www.magnetic-resonance.org/`, 2014.

[8] Magnetic resonance, a critical peer-reviewed introduction. In *European Magnetic Resonance Forum*, November 2014.

[9] Filler A. Magnetic resonance neurography and diffusion tensor imaging: origins, history, and clinical impact of the first 50,000 cases with an assessment of efficacy and utility in a prospective 5000-patient study group. *Neurosurgery*, 65(4 Suppl):29–43, 2009.

[10] J. Ashburner and K.J. Friston. Computing average shaped tissue probability templates. *NeuroImage*, 45:333–341, 2009.

[11] John Ashburner. Computational anatomy with the spm software. *Magnetic Resonance Imaging, 27 (8): 1163–74*, October 2001.

[12] John Ashburner and Karl J. Friston. Voxel-based morphometry-the methods. *Neuroimage, 11(6):805-821*, 2000.

[13] Asa Ben-Hur, David Horn, Hava Siegelmann, and Vladimir Vapnik. Support vector clustering. *Machine Learning Research*, (2):125–137, 2001.

[14] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, 1995.

[15] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer science+business media llc edition, 2006.

[16] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory – COLT*, page 144, 1992.

[17] H. Brodaty, M.M. Breteler, S.T. Dekosky, P. Dorenlot, L. Fratiglioni, and C. Hock. The world of dementia beyond 2020. (59(5):923-7), May 2011.

[18] A. Burns and S. Iliffe. Alzheimer's disease. *University of Manchester Psychiatry Research Group*, page BMJ 338: b158.doi:10.1136/bmj.b158. PMID 19196745., February 2009.

[19] Carlton Chu, Ai-Ling Hsu, Kun-Hsien Chou, Peter Bandettini, and ChingPo Lin. Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Alzheimer's Disease Neuroimaging Initiative*, 2011.

[20] C.Y.C. Chu. *Pattern recognition and machine learning for magnetic resonance images with kernel methods.* PhD thesis, Wellcome Trust centre for Neuroimaging. University College London, London., 2009.

[21] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning 20 (3): 273*, 1995.

[22] Corinna Cortes and V.N. Vapnik. Support vector networks. 1995.

[23] S.G. Costafreda, C. Chu, J. Ashburner, and C.H. Fu. Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One 4, e6353*, 2009.

[24] Nello Cristianini. Kernel methods for general pattern analysis. 2005.

[25] Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. *Cambridge University Press, The Edimburg Building, Cambridge CB*, 2005.

[26] H. A. David and Jason L. Gunnink. The paired t test under artificial pairing. *The American Statistician*, 51(1):9–12, 1997.

[27] Pierre A. Devijver and Josef Kittler. *Pattern Recognition: A Statistical Approach.* Prentice-Hall, 1982.

[28] Werner Dubitzky, Martin Granzow, and Daniel Berrar. Fundamentals of data mining in genomics and proteomics. *Springer Science and Business Media*, page 178, 2007.

[29] H.M. Duvernoy. *The Human Brain: Surface, Blood Supply, and Three Dimensional Sectional Anatomy.* Springer, second edition edition, 1999.

[30] Barbara Fadem. *High-Yield Behavioral Science (High-Yield Series).* Hagerstwon, MD: Lippincott Williams and Wilkins, isbn 0-7817-8258-9 edition, 2008.

[31] G.B. Frisoni, C. Testa, A. Zorzan, F. Sabattoli, A. Beltramello, H. Soininen, and M.P. Laakso. Detection of grey matter loss in mild alzheimer's disease with voxel based morphometry. *J. Neurol. Neurosurg. Psychiatry*, 73:657–664, 2002.

[32] G.B. Frisoni, C. Testa, A. Zorzan, F. Sabattoli, A. Beltramello, H. Soininen, and M.P. Laakso. Detection of grey matter loss in mild alzheimer's disease with voxel based morphometry. *J. Neurol. Neurosurg. Psychiatry*, 73:657–664, 2002.

[33] Chavhan G.B. *MRI Made Easy (for Beginners).* Jaypee Brothers Medical Publishers Pvt. Ltd, isbn:9351520471 edition, 2013.

[34] Seymour Geisser. *Predictive Inference.* Chapman and Hall. ISBN 0-412-03471-9, 1993.

[35] D. Getsios, K. Migliaccio-Walle, and J.J. Caro. Nice cost-effectiveness appraisal of cholinesterase inhibitors: was the right question posed? were the best tools used? *Pharmacoeconomics*, 25(12):997–1006, 2007.

[36] Richard Goering. Matlab edges closer to electronic design automation world. *EE Times*, 2004.

[37] Commissioning guide 2007. *NICE Memory assessment service for the early identification and care of people with dementia.* London: National Institute for Clinical Excellence, July 2011.

[38] Steve R. Gunn. Support vector machines for classification and regression. 1998.

[39] Steve R. Gunn. Sparse kernel methods. 2003.

[40] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classi cation using support vector machines. *Machine Learning*, (46):389–422, 2002.

[41] Isabelle Guyon and Andre Elissee. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, (3:1157-1182), 2003.

[42] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and L. Zadeh. Feature extraction. *Foundations and applications*, 2006.

[43] J. Hammon. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale. November 2013.

[44] Bertha Hidalgo and Melody Goodman. Multivariate or multivariable regression? *American journal of public health*, 103(1):39–40, 2013.

[45] Ovidiu Ivanciuc. Applications of support vector machines in chemistry. 2007.

[46] F.H. Joanneum. Cross-validation explained. *Institute for Genomics and Bioinformatics*, 2006.

[47] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (San Mateo, CA: Morgan Kaufmann)*, volume 2, page 1137–1143, 1995.

[48] Sasaki M., Ehara S., Nakasato T., Tamakawa Y., Kuboya Y., Sugisawa M., and Sato T. Mr of the shoulder with a 0.2-t permanent-magnet unit. *AJR Am J Roentgeno*, 154(4):777–8, April 1990.

[49] Richard Mankiewicz. The story of mathematics. *Princeton University Press*, page 158, 2004.

[50] Geoffrey J. McLachlan, Kim-Anh Do, and Christophe Ambroise. Analyzing microarray gene expression data. *Wiley*, 2004.

[51] Donald W. McRobbie. *MRI from picture to proton.* New York: Cambridge University Press, isbn 0-521-68384-x edition, 2007.

[52] Miguel González Mendoza. Aprendizaje estadístico, redes neuronales y support vector machines: Un enfoque global. *TESM CEM: Intelligent Transportation Systems Research Group. Universidad Veracruzana, Xalapa, México*, 2005.

[53] Tom M. Mitchell. *Machine Learning.* McGraw-Hill International Edition, 1997.

[54] Andrew W. Moore. Cross-validation for detecting and preventing overfitting. *Carnegie Mellon University.*

[55] J Graziella Orru, William Pettersson-Yeo, Andre F Marquand, Giuseppe Sartori, and Andrea Mechelli. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience and Biobehavioral Reviews, 36(4):1140-1152*, 2012.

[56] Penny, Friston, Ashburner, Kiebel, and Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 1st edition edition, 2006.

[57] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classi ers and fmri: a tutorial overview. *Neuroimage*, 45(1):199–209, 2009.

[58] Payam Refaeilzadeh, Lei Tang, and Huan Lu. k-fold cross-validation. *Arizona State University*, 2008.

[59] U. Salvolini and Scarabino T. *High Field Brain MRI*. Springer, isbn:3540317759 edition, 2006.

[60] G.M. Savva and C. Brayne. *Epidemiología y repercusión de la demencia. Manual de Enfermedad de Alzheimer y otras demencias*, 2011.

[61] J. Talairach and P. Tournoux. Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system - an approach to cerebral imaging. *Thieme Medical Publishers, New York*, 1988.

[62] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. Recent advances of large-scale linear classification. *Proc. IEEE 100 (9)*, 2012.

[63] Daisy Yuhas. `http://blogs.scientificamerican.com/`, June 2012.